UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot · your knowledge partner

Multivariate Nonlinear Time Series Analysis of Dynamic Process Systems

Gorden Takawadiyi Jemwa

Thesis submitted in partial fulfilment of the requirements for the degree Master of Science in Engineering (Extractive Metallurgical Engineering) in the Department of Chemical Engineering at the University of Stellenbosch

Study Leader: Professor C. Aldrich

April 2003

DECLARATION

I, the undersigned, hereby declare that the work contained in this thesis is my own original work and has not previously been submitted at any university for a degree in its entirety or in part.

Gorden Takawadiyi Jemwa, 17 February 2003

Abstract

Physical systems encountered in process engineering are invariably ill-defined, multivariate, and exhibit complex nonlinear dynamical behaviour. The increasing demands for better process efficiency and high product quality have led to the development and implementation of advanced control strategies in process plants. These modern control strategies are based on the use of a mathematical model defined for the process. Traditionally, linear models have been used to approximate the dynamics of processes whereas most processes are governed by nonlinear mechanisms. Since linear systems theory is well-established whereas nonlinear systems theory is not, recent developments in nonlinear dynamical systems theory present opportunities for improved approaches in modelling these process systems. It is now known that a nonlinear description of a process can be obtained from using time-delayed copies reconstructed from measurements taken from the process. Due to low signal to noise ratios associated with measured data it is logical to exploit redundant information in multivariate time signals taken from the systems in reconstructing the underlying dynamics.

This study investigated the extension of univariate nonlinear time series analysis to the situation where multivariate measurements are available. Using simulated data from a coupled continuously stirred tank reactor and measured data from a flotation process system, the comparative advantages of using multivariate and univariate state space reconstructions were investigated. With respect to detection of nonlinearity multivariate surrogate analysis were found to give potentially robust results because of preservation of cross-correlations among components in the surrogate data. Multivariate local linear models showed a deterministic structure in both small and large neighbourhood sizes whereas for scalar embeddings determinism was defined only in smaller neighbourhood sizes. Non-uniform multivariate embeddings gave local linear models that resembled models from a trivial reconstruction of the

original state space variables. With regard to global nonlinear modelling, multivariate embeddings gave models with better predictability irrespective of the model class used. Further improvements in the performance of models were obtained for multivariate non-uniform embeddings.

A relatively new statistical learning algorithm, the least-squares support vector machine (LSSVM), was evaluated using multilayer perceptrons (MLP) as a benchmark in modelling nonlinear time series using simulated and plant data. It was observed that in the absence of autocorrelations in the variables and sparse data LSSVMs performed better than MLPs. Simulation of trained models gave consistent results for the LSSVMs, which was not the case for MLPs. However, the computational costs incurred in training the LSSVM model was significantly higher than for MLPs. LSSVMs were found to be insensitive to dimensionality reduction methods whereas the performance of MLPs degraded with increasing complexity of the dimension reduction method. No relative merits were found for using complex subspace dimension reduction methods for the data used. No general conclusions could be drawn with respect to the relative superiority of one class of models method over the other.

Spatiotemporal structures are routinely observed in many chemical systems, such as reactive-diffusion and other pattern forming systems. We investigated the modelling of spatiotemporal time series using the coupled logistic map lattice as a case study. It was found that including both spatial and temporal information improved the performance of the fitted models. However, the superiority of spatiotemporal embeddings over individual time series was found to be defined for certain choices of the spatial and temporal embedding parameters.

Opsomming

Fisiese stelsels wat in prosesingenieurswese voorkom is dikwels nie goed gedefinieer nie, multiveranderlik en vertoon komplekse nie-lineêre gedrag. Toenemende vereistes vir hoë prosesdoeltreffendheid en produkgehalte het gelei tot die ontwikkeling en implementering van gevorderde beheerstrategieë vir prosesaanlegte. Hierdie morderne beheerstrategieë is gebaseer op die gebruik van wiskundige prosesmodelle. Lineêre modelle word gewoonlik ontwikkel, al is die onderliggende prosesmeganismes in die algemeen nie-lineêre, aangesien lineêre stetselteorie goed gevestig is, en nie-lineëre stelselteorie nie. Onlangse verwikkelinge in die teorie van nie-lineêredinamiese stelsels bied egter geleenthede vir verbeterde modellering van prosesstelsels. Dit is bekend dat 'n nie-lineêre beskrywing van 'n progses verkry kan word deur tydvertraagde kopieë van metings van die prosesse te rekonstrueer. Met die lae sein-tot-geraasverhoudings wat met gemete data geassosieer word, is dit logies om die oortollige informasie in meerveranderlike seine te benut tydens die rekonstruksie van die onderliggende prosesdinamika.

In die tesis is die uitbreiding van enkel-veranderlike nie-lineêre tydreeksontleding na meer-veranderlike stelsels ondersoek. Met data van twee aaneengeskakelde gesimuleerde geroerde tenkreaktore en werklike data van 'n flottasieproses, is die meriete van enkel- en meerveranderlike rekonstruksies van toestandruimtes ondersoek. Meerveranderlike surrogaatdata-ontleding het nie-liniariteite in die data op 'n meer robuuste wyse geïdentifiseer, a.g.v. die behoud van kruis-korrelasies in die komponente van die data. Meerveranderlike lokale lineêre modelle het 'n deterministiese struktuur in beide klein en groot naasliggende omgewings geïdentifiseer, terwyl enkelveranderlike metodes dit slegs vir klein naasliggende omgewings kon doen. Nie-uniforme meerveranderlike inbeddings het lokale lineêre modelle gegenereer wat soos globale modelle afkomstig van triviale rekonstruksies van die data gelyk het. M.b.t globale nie-lineêre modellering, het meerveranderlike inbedding deurgaans

beter modelle opgelewer. Verdere verbetering in die prestasie van modelle kon verkry word d.m.v. meerveranderlike nie-uniforme inbedding.

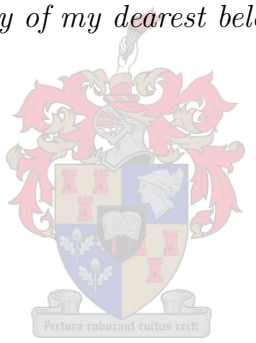
'n Relatief nuwe statistiese algoritme, die kleinste-kwadrate-steunvektormasjien (KKSVM) is geëvalueer teenoor multilaag-perseptrons (MLP) as 'n standaard vir die modellering van nie-lineêre tydreekse, deur gebruik te maak van gesimuleerde en werklike aanlegdata. Daar is gevind dat die KKSVM beter presteer het as die MLPs wanneer die opeenvolgende waarnemings swak gekorreleer en min was relatief tot die aantal veranderlikes. Die KKSVMs het beduidend langer geneem as die MLPs om te ontwikkel. Hulle was ook minder sensitief vir die metodes wat gevolg is om die dimensionaliteit van die data te verlaag, anders as die MLPs. Ook is gevind dat meer komplekse metodes tot die verlaging van die dimensionaliteit weinig nut gehad het. Geen algemene gevolgtrekkings kan egter gemaak word m.b.t die verskillende modelle nie.

Ruimtelik-temporale strukture word algemeen waargeneem in baie chemiese stelsels, soos reaktiewe diffusie e.a. patroonvormende sisteme. Die modellering van ruimtelik-temporale stelsels is bestudeer aan die hand van 'n gekoppelde logistiese projeksierooster. Insluiting van beide die ruimtelike en temporale inligting het tot beduidend beter modelle gelei, solank as wat dié inligting op die regte wyse ontsluit is.

Reflections on the way life was, the way life is, the way life might be – peaceful meanderings through the mysteries of life, the journeys where we touch days, approach new days with the experience these mysteries have given us, with the revelations that still lie before us WAITING TO BE DISCOVERED

– Roma Ryan

To my parents, Martin and Demetria...the love you have shown me knows no bounds. In memory of my dearest beloved aunt, Matilda Chinyani.



Contents

List of Tables

List of Figures

Acknowledgements

Achievements are rarely the sole act of one person; rather it is the culmination of efforts and commitment of different individuals into a whole that is greater than the sum of the constituent parts. Indeed, the same holds for the work embodied in this thesis.

My thanks go to the University of Stellenbosch's Department of Chemical Engineering for the opportunity to be involved in its research efforts.

I extend my most deep felt gratitude to Professor Chris Aldrich for introducing and navigating me through this wondrous and sometimes scary adventure in nonlinear dynamics and exploratory data analysis.

I thank nonlinear dynamics and machine learning researchers worldwide for their insightful contributions. In particular, Michael Small for his advice, comments, and software implementations; Thomas Schreiber for helping me understand the Information Theoretic concept; Rainer Hegger for his multivariate local linear predictor implementation; Ricardo-Carretero-González for his code for estimation of sub-system characteristic exponents for the coupled logistic map; SISTA team at Katholieke Universiteit van Leuven for their LS-SVMlab© software; the TISEAN© team at Institut für Physikalische und Theoretische Chemie, Universtät Frankfurt, for their nonlinear time series analysis software; the TSTOOL© team at Drittes Physikalisches Institut, Georg-August-Universität, Göttingen; my colleagues for the fun and explorations into the unknown; and Ndeke for introducing me to LaTeX 2_ε. JP Barnard for his QuickIdent© nonlinear time series implementation and useful comments.

A special thank you to Juliana Steyl; you have been such a kind-hearted and helpful person ever since I stepped my feet into Stellenbosch. I would not have found it easier to fit into the rhythm of life in Stellenbosch without your help. Thank you very much.

A special mention of the personnel in the Department of Chemical Engineering; you are such a jolly good people to work with.
Thank you to my colleagues at the US's WritingLab for their advice and company.
I extend a special thank you to the van der Merwe family for being such wonderful and admirable hosts; you are family to me.
Warm hugs to my mom, dad, brothers, and sisters for their understanding and faith in me. I love you too much.
My greatest thanks to The Almighty God for the life He blessed me with.

Nomenclature

Notation conventions used in this thesis. Note some symbols may have different usage in some contexts.

x	a real-valued scalar quantity
$\mathbf{x}$	vector quantity
$\phi : M \rightarrow \Gamma$	a mapping function from a coordinate space M to a coordinate space Γ
$\mathbb{R}^n$	a vector space of n-dimensions
d_e	embedding dimension
d	state space dimension
T	time delay or delay lag
C_T	the autocorrelation function
I_T	average mutual information
$p(\cdot), p(\cdot, \ldots, \cdot)$	probability and joint probability distributions functions
μ	first-order moment or mean
σ	standard deviation/noise
$\mathbf{X}$	a multidimensional vector matrix
$\mathcal{A}$	an attractor of a dissipative system
ρ	an ergodic probability measure of a dynamical system that measures the frequency with which different points are traversed by a trajectory
$h(\rho)$	Kolmogorov-Sinai entropy
ε	neighbourhood size

D_F	capacity dimension
C_q	Generalized correlation sum
D_q	Generalized $q-$order Renyi dimensions
D_1	Information dimension
D_2	Correlation dimension
λ_i	Lyapunov exponent
λ	Regularization constant (modelling)
$\mathbf{J}$	Jacobian square matrix
$R_{emp}[f]$	Empirical risk function
$\ell(\cdot)$	loss function
$R[f]$	expected error or risk function
$\mathbf{w}$	normal vector of a hyperplane (neural networks and support vector machines)
ξ_i	the slack variable for pattern $\mathbf{x}_i$
$(\mathbf{x} \cdot \mathbf{y})$	scalar product between $\mathbf{x}$ and $\mathbf{y}$)
α_i, α_i^*	Langrange multiplier/Expansion coefficient for $\mathbf{w}$
ν_i, ν_i^*	Langrange multiplier/Expansion coefficient for the slack variable
$K(\cdot, \cdot)$	scalar product in feature space
$\| \cdot \|_p$	the ℓ_p norm, $p \in [1, \infty]$
$d_c(\varepsilon_0)$	correlation dimension estimate as a function of scale ε_0
$L(\cdot, \ldots, \cdot)$	Lagrangian function
CML	coupled map lattice
CSTR	continuously stirred tank reactor
ICA	independent component analysis
LSSVM	least squares support vector machines
MLP	multilayer perceptron
MSE	mean square error
PCA	principal component analysis
W	whitening

Chapter 1

Introduction

"Nothing in nature is random...A thing appears random through the incompleteness of our knowledge" – Spinoza, *Ethics I*

Exordium

In philosophy, the reason for human existence can be seen from two antithetical poles; time existence and eternal existence. A similar dialectical tension is observed within the dynamical systems discipline in the quest to understand the behaviour of physical systems. One perspective sees nature as a completely *deductive* system whose temporal evolution is according to some deterministic laws and, therefore, making absolute predictability possible. The apogee of this view is the classical Newtonian-Laplacian mechanics modelling approach. The second conflicting perspective views nature as an *inferential* system in which only statistical regularity is possible. In other words, the best we can hope for is to use estimation and approximation techniques on finite data observed on the system and project some structure consistent with observed distribution properties of the data. This is the approach taken in classical time series analysis where all events are assumed random and

occur according to some probability distribution.

There is yet no universal agreement on what *complexity* in nature is. Notwithstanding this, what constitutes complexity can be understood by consideration of the antithesis that exists in the scientific description of physical systems. Complexity attempts to offer a plausible understanding of the behaviour of physical systems by offering itself as an intermediate alternative in the otherwise irresoluble conflict between the two contentions of order and randomness. Thus, in the study of complexity we seek answers to the following questions (Crutchfield *et al.*, 1992):

- Are the two opposed views of scientific description of order and randomness not incomplete projections of a complex nature?
- Is nature too intricate and too detailed to remain completely irresolvable at any single time and with finite knowledge?
- Which physical system's state is not deterministic (within some degree of accuracy) with any finite measurement?

Modern approaches in dynamical systems theory attempt to solve the inverse modelling problem by learning the behaviour of the system using information generated by the physical system. This is achieved by, for example, reconstructing the state space of a system using the generated signals. This is the nonlinear equivalent of system identification techniques widely used in linear modelling. The validity of this approach was shown empirically by Packard *et al.* (1980). A mathematical justification was proven independently by Takens (Abarbanel, 1996; Kantz and Schreiber, 1997; Sauer *et al.*, 1990) and is embodied in *Takens' embedding theorem*, on which nonlinear time series analysis is firmly founded.

1.1 Complexity in Chemical Process Systems

Many chemical and metallurgical processes are characterized by highly nonlinear and complex dynamics, with long time constants and significant delays. The presence of nonlinearities gives rise to structural kinetic instabilities. Lee and Chang (1996) refer to a number of chemical and biochemical systems that have been shown to exhibit chaotic dynamics, a specific nonlinear behaviour of interest. Examples of such systems include chaos in the dynamics of solution polymerization of vinyl acetate in a full-scale continuously stirred tank reactor; parallel cubic autocatalytic reactor; a forced exothermic chemical reactor; a fluidized bed catalytic reactor with consecutive exothermic chemical reactions effected by changes in system parameters; nonlinear chaotic behaviour in an industrial process involving oxidation of *p*-xylene to terephthalic acid; etc.

The origin of complex behaviour can be understood if one considers the following: Typical process systems involve unit operations, with various complicated reactions, and heat/mass transfer processes occurring within each unit or subsystem. Integration of unit operations into a larger system invariably results in an upstream unit operation acting as a driving force for downstream operations. At a microscopic level, individual reaction sites perturb nearby sites that, in turn, induce a similar effect on their neighbouring sites. A visualization of these interactions is very difficult. However, the effects eventually manifest themselves globally in a seemingly irregular way.

The analysis of such systems using purely statistical modelling approaches is restrictive as causal effects other than external random inputs (noise) or time delays can be identified and possibly be isolated. The opposite extreme alternative approach using Newtonian differential equations is also inadequate. Not only is the process of deriving the underlying equations of motion cumbersome, it is also not guaranteed that such equations can be found and, if they exist, whether they are

tractable (Rapp *et al.*, 1999). The utility of complexity as an arbiter between the two approaches becomes handy, as alluded to earlier. It then is necessary to formulate, develop and formalize different, possibly novel, approaches in the simulation, modelling, optimization, and control of chemical and metallurgical processes.

Useful models that explain observed complex dynamical behaviour of many chemical reactions and reactors have been developed in previous studies, for example Gray and Scott (1983, 1984) and Jorgensen and Aris (1983). Analysis of these mathematical models plays an important role in the understanding and control of practical reactors. If adapted to physical realities the models allow for the study of possible anomalous behaviour and guidelines on how to avoid or exploit this behaviour, depending on whether the behaviour is desirable or not. In particular, discovery of the existence of chaotic phenomena in physical systems has broadened the space over which the behaviour of reactive systems can be explored with a view towards control for better process yield and selectivity (especially in cases where undesirable competing reactions can occur). For examples, Hoffman and Schadlich (1986) and Silverton *et al.* (1986) show that operating around oscillatory trajectories can improve process performance.

In the following some of these examples are discussed in more detail.

1.1.1 The continuous stirred tank reactor

The continuously stirred tank reactor (CSTR) is a common model reactor used in process modelling and simulation tasks to approximate reactor dynamics. Conceptually, a CSTR is quite simple – a well mixed tank that facilitates contact amongst reactants. A continuous inflow and outflow of reactants and products respectively in the reactor is assumed. The feed assumes a uniform composition throughout the reactor and, therefore, the outflow stream has same uniform composition as in the tank. In spite of its simplicity, mixing patterns in a CSTR can be very complex.

Hopf bifurcation, multiple steady states and other pathological reaction behaviours are routinely encountered in both isothermal and non-isothermal process systems.

In non-isothermal CSTR systems, complex dynamical behaviour for reactions are induced by, for example, thermal feedback. Here temperature variations affect the rate of reaction across a parametric range associated with both simple and complex behaviours. Mankin and Hudson (1984) investigated the effect of perturbing a non-isothermal reaction with an Arrhenius temperature dependence and an oscillatory basic state. The authors showed that by varying the amplitude of the coolant temperature the following sequence of dynamical behaviour is observed: quasi-periodic → periodic → bifurcation → chaos. The catalytic exothermic consecutive reaction network

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C \tag{1.1}$$

with one exothermic and one endothermic reaction is known to exhibit complex dynamics with various patterns of multiplicities of the steady states. Jorgensen and Aris (1983) investigated this system and showed evidence of complicated behaviour in regions of parameter space for which there is only one unstable steady state. Understanding and identifying such dynamical behaviour allows for an effective process control strategy that maximizes the process yield of the desired product. The dynamic behaviour of a system represented in equation (1.1) but occuring in a bubbling fluidized-bed reactor was investigated by Elnashie *et al.* (1995), where the intermediate species was the desired product. The challenge was to operate the reactor at the middle unstable steady state that maximized yield of species B. Practical examples in which such situations are encountered include

1. Gas-phase catalytic oxidation of hydrocarbons in the petrochemical industry. An example is the partial oxidation of *o*-xylene to phthalic anhydride (Elnashie *et al.*, 1995).

2. Oxidative dehydrogenation reactions such as the partial oxidative dehydrogenation of butene to butadiene (Wagialla *et al.*, 1991).

A feedback control or cooling system strategy can be used to maximize the yield of the intermediate product. This entails development of a suitable model that spans the dynamical space associated with system.

Chemical feedback is the other route through which complex dynamical behaviour occurs in non-isothermal CSTR process systems. In this case a product of reaction increases the rate of reaction and, consequently, its own rate of production. This is known as *autocatalysis*. The generalized prototype reaction mechanism for an autocatalytic reaction can be expressed as,

$$\begin{array}{rcll} A + mB & \rightarrow & (m+1)B, & r_A = -k_1 AB^m \\ B & \rightarrow & C, & r_C = k_2 B \end{array} \tag{1.2}$$

where r_i is the rate of reaction with respect to species i, m is an integer value, and $\{k_1, k_2\}$ are rate of reaction constant. For $m = 1$ and $m = 2$ one obtains quadratic autocatalysis and cubic autocatalysis respectively (Gray and Scott, 1983, 1984). Using an algebraic analysis, Gray and Scott showed that the system exhibited various patterns of exotic behaviour including multistability, hysteresis, etc. An interesting observation was that analogies could be drawn between isothermal systems and non-isothermal reactions. However, unlike the non-isothermal case, the system in equation (1.2) is not capable of displaying chaotic behaviour. Fortunately, Lynch (1992a) showed that by introducing a second autocatalytic step,

$$D + mB \rightarrow (m+1)B, \qquad r_D = -k_3 C_D C_B^m \tag{1.3}$$

the system can be described by three independent ordinary differential equations (a necessary but not sufficient condition for chaos) making it is possible to observe higher levels of complex behaviour including chaos. Models for the non-isothermal case are multi-parametric and stiff whilst those for isothermal systems are not.

Hence, use of the isothermal model allows for tractable analytical study of complex behaviour in reactive systems.

The CSTR is evidently a rich system that can be explored for the analytical study of complex behaviour in chemical process systems without necessarily observing a real system, a costly exercise vulnerable to the vissicitudes of experimentation.

1.1.2 Electrochemical reaction dynamics

Electrochemical reactions display variegated dynamical behaviour. Hudson and Tsotsis (1994) give an accessible review on the status of research into the dynamics of electrochemical reactions. Variation of a parameter in an electrochemical reactor parameter, such as voltage or current, results in change in dynamical behaviour. These observed behaviours include bi-stability, oscillatory, period-doubling bifurcation, quasi-periodicity and chaos. Spatiotemporal patterns develop due to coupling of nonlinear reaction sites by mass transfer of ion-pairs through the electric field in the electrolyte for example. In some systems it has been observed that increasing the area of the electrode surface results in a change in the complexity of the system as measured by the dimensionality estimates (Green *et al.*, 2000).

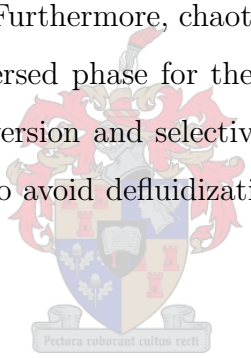
Electrochemical reactions are of practical importance in many areas. Typical examples include corrosion control in metals; electro-deposition processes in hydrometallurgical refining operations; and the direct electrochemical oxidation of various electro-organic compounds in fuel cells.

1.1.3 Multiphase reactors

Multiphase reactors involve the interaction of at least two phases, such as gas-liquid, liquid-liquid, gas-solid, liquid-solid, and gas-liquid-solid. They are used primarily to alter mass transfer coefficients in reacting systems. Some of the most commonly used multiphase reactors are spray towers, bubble columns, kilns, fluidized beds,

etc. It is now accepted that certain multiphase reactors exhibit chaotic dynamical behaviour. For example, research on industrial units has indicated that fluid catalytic cracking units and fluidized-bed polyethylene units show complex static and dynamic bifurcation (Elnashie *et al.*, 1995).

Chaotic behaviour confronts the design engineer with difficulties in the design and scale-up of multiphase reactors. Van der Bleek *et al.* (2000) proposed that identification of chaotic behaviour in these reactors offers possibilities for better characterization of the hydrodynamical profiles and improvements on currently used design scale-up laws. Using Kolgoromov's entropy measure, they show that a chaos-related scale-up law can be derived that better captures hydrodynamic regimes encountered in multiphase reactors. Furthermore, chaotic behaviour can be exploited to control the pattern of the dispersed phase for the purposes of improving mass transfer properties (improved conversion and selectivity). Potential application is in developing prognostic systems to avoid defluidization of the packed bed reactor through agglomeration.



1.1.4 Electrochemical noise analysis

Electrochemical noise (EN) is used to detect general and localized (stochastic) corrosion rates in equipment (Holcomb *et al.*, 2002). EN measurements are based on fluctuations in electrochemical potential and corrosion current that occur during electro dissolution. Relationships can be defined between the measured potential and the driving force of the reaction (Gibb's free energy, a thermodynamic quantity), and similarly between the corrosion current and rate of reaction (a kinetic quantity). Random surface electrochemical events occuring on a corroding metal generate noise in the overall potential and current signals. Each type of corrosion is associated with a certain structure in the signal noise. Analysis of the signals can be used in modelling the type and severity of the electro-dissolution process.

Traditionally, characterization of the electrochemical response of systems undergoing localized corrosion has been done by classical time-domain, statistical, and frequency-domain approaches (Cottis *et al.*, 2001). The application of chaos theory concepts in EN analysis has been attempted and initial results have generated a lot of interest for further research in this direction.

1.2 Problem Formulation

Modelling plays a pivotal role in efforts geared towards improving industrial operations and process performance. Three major purposes of modelling can be distinguished;

1. Gain a better understanding or knowledge of the process, and of the interaction of the process with its environment;
2. Direct the knowledge thus obtained to predict and anticipate future behaviour of the system;
3. Manipulate the process to direct the system behaviour towards a desirable region of the state space.

There are two approaches used in deriving models:

- **Fundamental modelling** – A mathematical description of the process is derived from consideration of the physical laws governing the system; and
- **Empirical modelling or System Identification** – A mathematical model is obtained based on measurement data (input and/or output signals) taken from the system to describe the dynamics of the system. The intention here is not to describe the underlying physical processes that are responsible for the observed behaviour.

The modelling activity is related to the purpose to which the model is intended, *v*iz-a-viz, analysis, prediction, or control. The approach taken here is to provide a model primarily to serve as a basis in robust control design based on system identification techniques, using tools and concepts inspired by nonlinear dynamical systems theory. Traditionally, process control has been achieved with *Proportional-Integral-Derivative* (PID) design rules. PID-based controllers are still the most widely used controllers in many applications due to their simplicity. However, they are often unsatisfactory for complex or multivariable process and in cases where high performance restrictions are imposed on the controlled process. This limitation led to the development and growth of model-based control design rules. These are based on the realization that it is easier to construct an approximator to the system behaviour by fitting a model than by "tuning" PID parameters. Model-based control design methods make use of a state space model of the target system, and the controller is calculated according to specified criterion under the assumption of the *certainty-equivalence* principle. *Robust* control design and analysis have since evolved that take into consideration that the model is an approximate description to the controlled process. Hence, in addition to the nominal model, the error bounds on the model need to specified.

To put the perspective of where the work herewith fits in the broader framework, consider a typical process control depicted in Figure 1.1(a). The objective is to restrict the variation of a process response variable within the set points by using information from the past to forecast future behaviour. Variations between model ($\tilde{\mathbf{p}}$) output and actual process ($\mathbf{p}$) response are traced in time. An error ($\mathbf{e}$) calculated from future inputs ($\mathbf{r}$) and past model-process discrepancy ($\mathbf{y} - \mathbf{y}_m$) is fed into the controller ($\mathbf{q}_c$), which adjusts the inputs into the process such that deviation from set points are minimized. This is important in maintaining a consistent product quality in, for instance, metal refining processes. For example, in

nickel refining it is important to restrict levels of lead within a certain range else the resulting product is unusable.

Advanced control systems based on Model-Based Predictive Control (MPC) are now a common feature in state-of-the-art process plants. In such systems an optimally and explicitly defined process model ($\tilde{\mathbf{p}}$) is used in the control algorithms to predict the future behaviour of a plant. The most important task in MPC technology is the exact parameterization of the mathematical model of the process. Not surprisingly, roughly $70 - 80\%$ of efforts in MPC are focused in this area. The model allows the controller to deal with an almost exact replica of the real process dynamics, resulting in a much better control mechanism. Two features distinguish MPC technology from other process control techniques (Lazar and Pastravanu, 2002);

- The constraints with respect to input and output signals are directly considered in the control calculation, resulting in tighter control and improved controller reliability.

- MPC algorithms consider plant behaviour over a future horizon in time. Thus, the effects of both feed-forward and delay feedback disturbances can be anticipated and eliminated. This permits the controller to drive the process output more closely to the reference trajectory.

Most of the models being employed in MPC approaches are based on linear dynamical rules. Linear models are sufficiently accurate for processes with periodic oscillatory behaviour. However, it is accepted that most processes are usually best described by complex nonlinear equations. Furthermore, chaos theory has established that perfectly deterministic systems exhibit similar behaviour that is associated with random disturbances in linear modelling (Eckmann and Ruelle, 1985). Therefore, there is an increasing need to develop accurate nonlinear models.

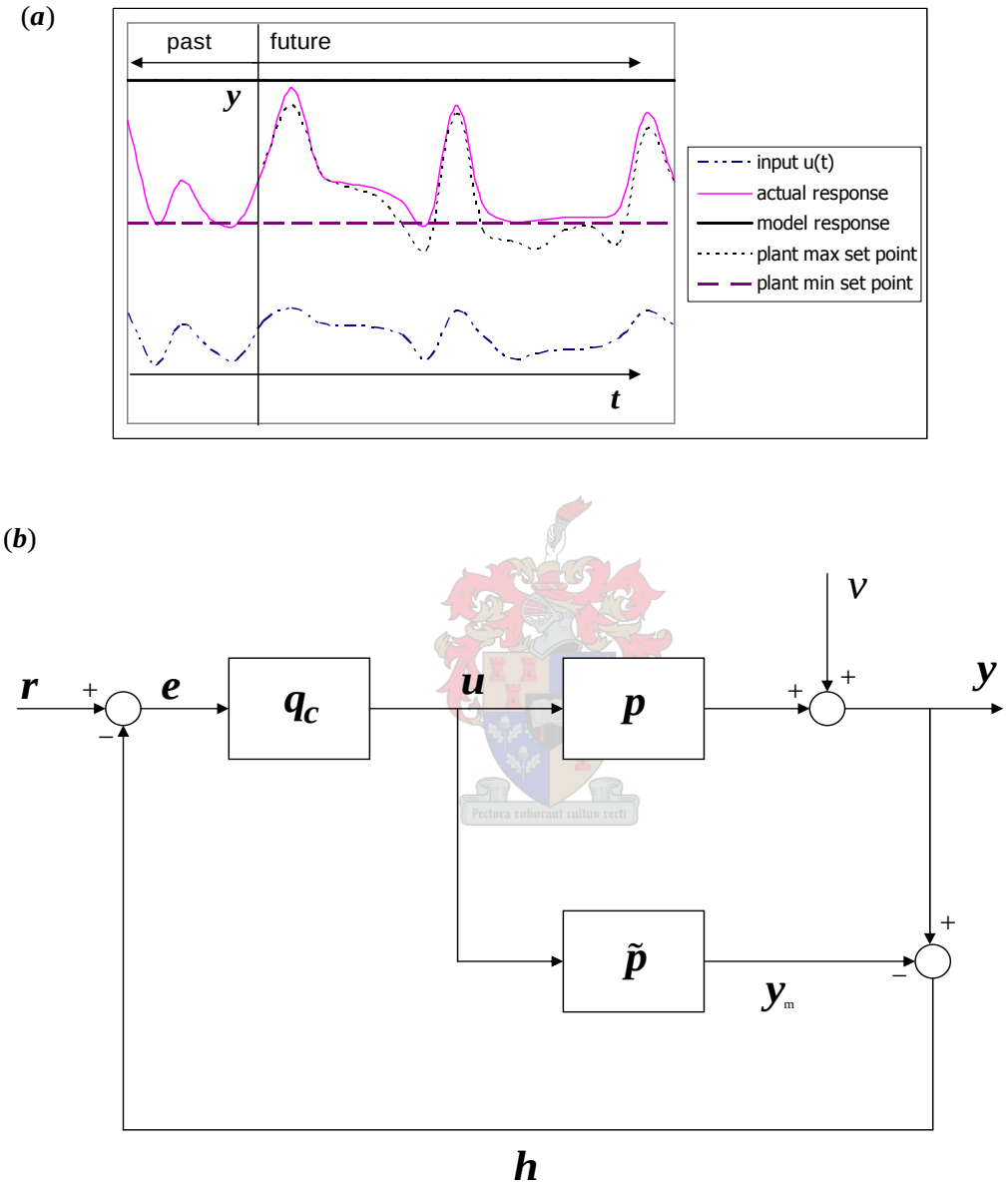


Figure 1.1: System identification in process control (a) The objectives in predictive control; (b) Control relevant identification block diagram.

However, the task of obtaining a robust model is considerably more difficult for nonlinear processes. Developments made in the last two decades within nonlinear dynamical systems field offer improved approaches for model development.

The concepts and techniques of nonlinear modelling are now well-formalized for univariate time series (Abarbanel, 1996; Kantz and Schreiber, 1997), although it should be added that these approaches are by no means well-established in the industrial process engineering community. Moreover, multivariate nonlinear time series analysis pose a special challenge because of their size, presence of high correlations between variables, and the low content of information in any single variable (low signal-to-noise ratios). The recent and ongoing advances being made in computational processing capacity and a reasonably sound theoretical framework provide a platform that allows for the redundant information in the multivariables to be exploited. As far as process industries are concerned these issues are very important since most process systems are represented by multivariate time series.

Therefore, this thesis emphasizes the fundamental analysis of multivariate systems, with the following specific objectives:

- Extension of state space reconstruction methods (both uniform and non-uniform) to multivariate time series. In particular, the effect of embedding strategies on model quality will be investigated, the theory being that in multivariate time series the correlation between variables (redundancy) can be exploited to generate better embeddings than possible when the variables are embedded individually. This concept has received little attention in literature to date.

- Comparison of the effect of principal component analysis and independent component analysis methods for subspace dimensionality reduction on modelling. This objective follows from the above-mentioned and has potentially important implications when time series are non-Gaussian.

- Comparison of characterization of nonlinear dynamical systems with different invariant quantities. Considerable progress has recently been made with respect to the characterization of dynamic systems from observed data. However, little has been published with respect to the relative merits of different criteria for characterizing time series.

- Handling of spatiotemporal time series as a special case of high-dimensional systems. These systems require excessive computational resources to analyze and have therefore received very little attention to date despite their obvious importance in chemical reaction engineering and other areas of process engineering.

- Finally, evaluation of a relatively new learning methodology based in statistical learning theory, namely support vector machines, in the modelling of high-dimensional data using multilayer perceptron networks as a benchmark. There are some indications that support vector machines are better suited to deal with sparse high-dimensional data, but this has not been established comprehensively as yet.

1.3 Outline of Thesis

The thesis is organized as follows;

- In Chapter 2 the concepts of nonlinear times series analysis inspired by chaos theory are reviewed. State space reconstruction techniques are discussed, as well as characterization of physical systems using three ergodic dynamical invariants, i.e. dimensions (effective degrees of freedom), entropy (rate of information production), and Lyapunov or characteristic exponents (sensitivity

to infinitesimal changes in initial conditions). Principal and independent component analysis methods for subspace dimension reduction are also reviewed.

- Chapter 3 looks at the inverse modelling problem of fitting equations of motion governing the dynamical evolution of physical systems using observed measurements taken from a physical system. The ideas underlying the use of multilayer perceptrons and especially support vector machines are treated in some detail. A variation of support vector machines called least-squares support vector machines is also considered.

- Chapter 4 investigates the application of the above-mentioned methodologies using a coupled continuously stirred tank reactor model as a case study. The performances of multilayer perceptron neural networks and support vector machines are compared.

- In Chapter 5 data from a flotation plant are analyzed to illustrate the practical implementation of nonlinear time series analysis using possibly multivariate time series.

- Chapter 6 investigates reconstruction and prediction of spatially extended systems, a special class of multivariate time series. The coupled map lattice concept is used to illustrate the advantages of including both spatial and temporal information in state space reconstruction.

- Chapter 7 summarizes the main conclusions from the study and recommendations for further work.

Chapter 2

Nonlinear Time Series Analysis

"Man is the only animal capable of conscious evolution; he invents tools."

– Alfred Russel Wallace

In this work, by *time series* we understand a time-ordered sequence of observations taken from a physical system at regular or irregular sampling intervals. The theoretical concepts of low-dimensional determinism have revolutionized approaches in time series analysis (Abarbanel, 1996; Abarbanel *et al.*, 1993; Eckmann and Ruelle, 1985; Kantz and Schreiber, 1997). It is now acknowledged that the occurrence of irregular and complicated behaviour that is seemingly induced by the action of external random perturbations can be explained using the theory of deterministic chaos. Previously, irregularity in a time series had been explained by traditional linear stochastic models, which assume that such signals are projections of a superposition of external random influences on otherwise linear dynamical rules. The linear stochastic approach has been explored comprehensively and extensive results can be found, for example, in the celebrated work of Box and Jenkins (1976). Tong (1990) gives a nonlinear time series analysis treatment using a stochastic approach.

Low-dimensional determinism provides an additional alternative set of mathematical tools for the characterization of irregular time series data generated by

complex phenomena. The fundamental properties of chaotic behaviour have been observed in simulated and physical experiments (Eckmann and Ruelle, 1985). The approach has also been shown to be useful even in instances where the data is not necessarily deterministic but contains patterns that cannot be explained adequately using linear techniques, for example, financial markets data. The theoretical development and tools of nonlinear time series are still immature compared to the alternative linear stochastic approach. This in itself is not a defect but rather a limitation that is being overcome as research and developments continue in the field.

A framework for the analysis of nonlinear dynamical systems is now well established (Abarbanel *et al.*, 1993). The fundamental concepts of chaotic systems (i.e., dimensions, sensitivity to changes in initial conditions, and entropy or production of information) provide tools for a systematic investigation and knowledge of dynamical phenomena. Figure 2.1 shows a broad hierarchical structure of the various issues dealt with in nonlinear time series analysis. As depicted in the illustration, nonlinear time series analysis is a continuously evolving process of discovery that projects a certain system behaviour, then challenges that assumption and, if necessary, reconfigures our perceived understanding on the basis of failure of the initial assumption or new information. (Note the representation is not necessarily complete but only highlights issues most pertinent to this work.)

2.1 State Space Reconstruction

The theoretical framework for nonlinear time series motivated by nonlinear dynamical systems theory is provided by Takens' *embedding theorem* (Takens, 1981)[1] and the *prevalence* extension (Sauer *et al.*, 1990). Essentially, the embedding theorems say that given a time series of some observable x_t, it is possible to reconstruct the

[1] cited in Abarbanel (1996) and Kantz and Schreiber (1997)

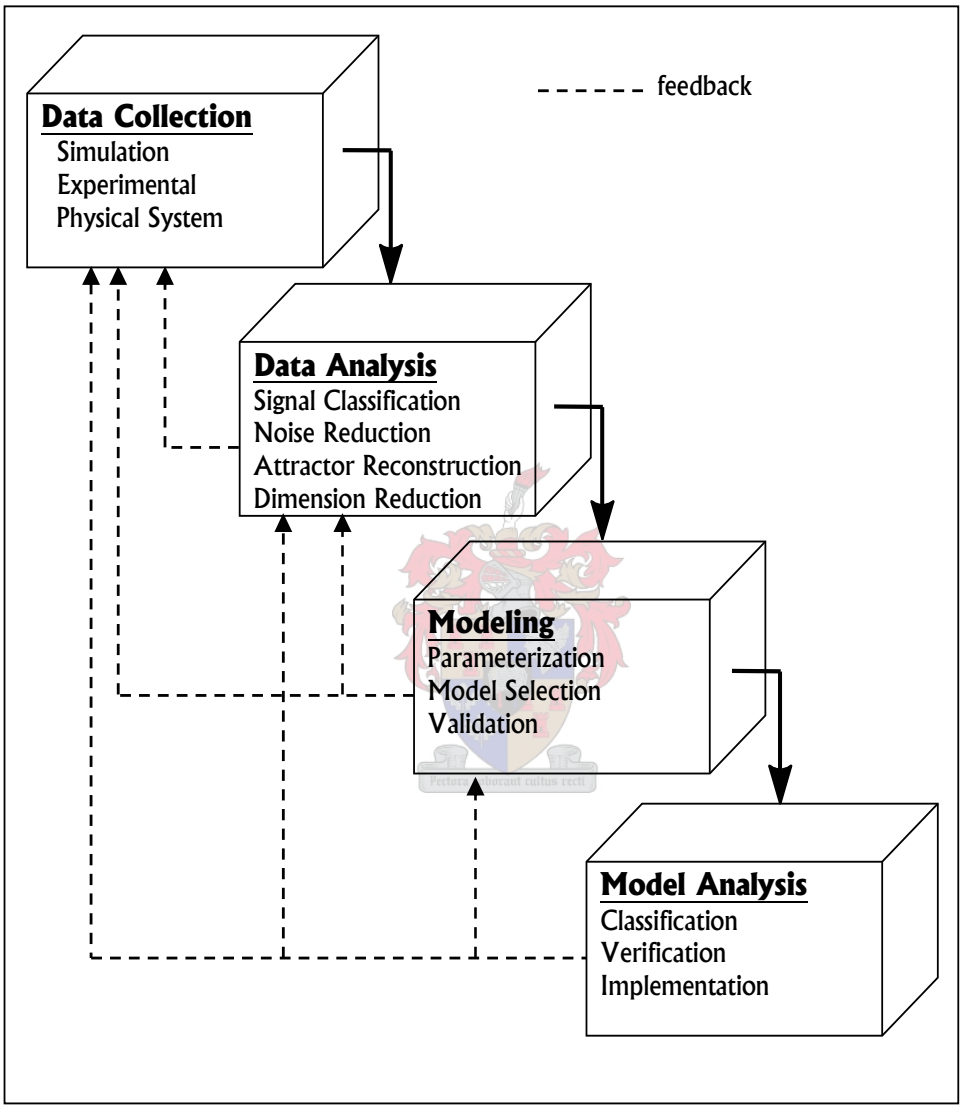


Figure 2.1: A hierarchical explosion of the different processes involved in nonlinear time series analysis and its cyclical nature.

state space of the dynamics generating the observable that is *diffeomorphically*[2] equivalent to the original state space $\mathbb{R}^d$. The reconstructed state space is called the *embedding space*. Assuming the system is deterministic[3], an observed (discrete or continuous) time series x_t on the system can therefore be used to reconstruct the state of the system at some given time.

The time evolution of a deterministic finite-dimensional dynamical system with state $\mathbf{x}$ in some manifold $M \subset \mathbb{R}^d$ is given by a map $\varphi^t : M \to M$ such that

$$\mathbf{x}_t = \varphi^t(\mathbf{x}_{t_0}) \tag{2.1}$$

where $\mathbf{x}_{t_0}$ and $\mathbf{x}_t$ are the state at time t_0 and t respectively. If h is some measurement function, $h : M \to \mathbb{R}^D$, the observed time series x_t is related to the states of the dynamical system by

$$x_t = h(\mathbf{x}_t), \qquad x_t \in \mathbb{R}^D \tag{2.2}$$

The embedding theorems assert that given only a *noise-free* and *infinite* scalar[4] time series generated by a nonlinear system, it is possible to reconstruct a diffeomorphically equivalent state space of the original state space by delay coordinates, provided the embedding dimension is sufficiently large enough. More formally;

Theorem 2.1 (Takens) *Let $M \subset \mathbb{R}^d$ be a smooth C^2 compact manifold that constitute the true state space of the dynamical system, $\varphi : M \to M$ is the corresponding flow ($\mathbf{x}_t = \varphi^t(\mathbf{x}_{t_0})$) and $h : M \to \mathbb{R}$ a measurement function. The mapping $\Phi : M \to \mathbb{R}^{d_e}$ defined by*

$$\Phi(\mathbf{x}) = \left\{ h(\varphi^{-d_e}(\mathbf{x})), h(\varphi^{-(d_e-1)}(\mathbf{x})), \ldots, h(\varphi^{-1}(\mathbf{x})) \right\} \tag{2.3}$$

[2] A *diffeomorphism* is a differentiable function whose inverse is also differentiable

[3] A system is called *deterministic* if specifying the current state completely defines the time behaviour of all future states.

[4] $D = 1$ in equation (2.2)

is an embedding for $d_e \geq 2d+1$ *under suitable the smoothness and genericity assumptions, that is,* $\Phi : M \to \Phi(M) \subset \mathbb{R}^{d_e}$ *is a* C^2*–diffeomorphism*

Restating in simpler terms, observed measurements are real-valued projections of unknown nonlinear combinations of the underlying state variables of the system and, therefore, completely retain all information of the state variables. Although Takens original formulation requires $d_e \geq 2d+1$ for a proper reconstruction Sauer *et al.* (1990) showed that any value $d_e > d$ can be used as long as the genericity conditions are satisfied. The requirement for diffeomorphic equivalence becomes apparent since typically both φ and h are unknown and, therefore, it is futile to attempt to reconstruct the original state space. The state space reconstruction problem is depicted in Figure 2.2.

Reconstruction of the state space implicitly assumes that the past (and future) observed measurements of a time series contain information about the (unobserved) state variables that can be used to define a state at the present time. This is typically done using delay coordinates. Assuming a *predictive* reconstruction, the d_e-dimensional delay coordinate vector at some time t is defined by

$$\mathbf{x}_t = (x_t, x_{t-T}, \ldots, x_{t-(d_e-1)T})' \tag{2.4}$$

The time separation between coordinates is called the *lag or delay time* T. Defining the delay reconstruction map Φ that maps the original states of a d-dimensional dynamical system into the embedded d_e-dimensional delay vectors as,

$$\Phi(x) = (h(\varphi(\mathbf{x})), h(\varphi^T(\mathbf{x})), \ldots, h(\varphi^{(d_e-1)T}(\mathbf{x}))) \tag{2.5}$$

then Φ is a smooth, one-to-one coordinate transformation with a smooth inverse, or *embedding*, when $d_e > d$. If Φ is an embedding then a smooth dynamics F is induced in the reconstructed phase space:

$$F^t(\mathbf{x}) = \Phi \circ \varphi^t \circ \Phi^{-1}(\mathbf{x}) \tag{2.6}$$

In other words, except for a coordinate change Φ, F and φ are similar. Therefore, all coordinate independent properties of F and φ are identical. Thus, geometrical invariants such as the eigenvalues of fixed and periodic points, generalized dimensions, Lyapunov exponents, and other topological features of the original state space are preserved in the reconstructed space. Thus, one can study the dynamical behaviour in the reconstructed state space instead of the true state space. An example of the reconstruction of phase space using embedding is shown in the appendices.

2.1.1 Choice of optimal time delay

The embedding theorems do not address the issue of the selection of an optimal lag since an infinite amount of quality signals is assumed to be available. In practice, however, the minimum embedding dimension is dependent on the choice of the time delay T. A good choice of T may decrease the minimum embedding dimension required in attractor reconstruction. Heuristic criteria have been developed for the choice of the time delay. Often used criteria select a T such that the value of the signal at time $t_0+n\tau_s$, x_n is *independent* or *uncorrelated* of the value of the signal at a later time $t_0+(n+T)\tau_s$, x_{n+T}, where τ_s is the sample interval. Such a selected T results in delay vector elements to be as independent as possible but still remaining connected to each other. In other words, such a time delay allows the effect of the (other) unobserved dynamical variables to be reflected in the value of x_{n+T}. Two common methods for choosing such a T are the *autocorrelation function* and *average mutual information*. The basis of each of these approaches is heuristic and it is not guaranteed for the values obtained using the different approaches to converge.

(a) ***Autocorrelation Function***

The autocorrelation function measures the expectation of observing the x_{n+T} at a time T later when x_n is observed. It is a second-order moment function

and is given by

$$C_T = \frac{\sum_{i=1}^{N}(x_n - \bar{x})(x_{n+T} - \bar{x})}{\sigma^2} \tag{2.7}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N}(x_n)$ and $\sigma^2 = \frac{1}{N-1}\sum_{i}^{N}(x_{i=1} - \bar{x})^2$ are the mean and variance of the data signal respectively. The autocorrelation function of deterministic systems decays exponentially with increasing lag. Using a T at which C_T attains its first zero makes the coordinates linearly uncorrelated and, hence, a good approximation for the optimal T. Some authors note that in some cases such a criterion completely removes any connection between coordinates making proper reconstruction impossible (Kantz and Schreiber, 1997). Instead they suggest a time delay when C_T first decays to $1/e$.

(b) ***Mutual Information***

The average mutual information uses ideas from Information Theory to define an optimal time delay T. The average mutual information is the amount of information (in bits) learned by measurements of x_n through measurements of x_{n+T} and is given by

$$I_T = \sum_{x_n, x_{n+T}} p(x_n, x_{n+T}) \log_2 \frac{p(x_n, x_{n+T})}{p(x_n)p(x_{n+T})} \tag{2.8}$$

where $p(\cdot)$ and $p(\cdot,\cdot)$ are the probability and joint probability functions respectively.

Fraser and Swinney (1986) suggested that the first T where the first minimum of I_T occurs ensures attractor unfolding in a time delay embedding. I_T is the nonlinear equivalent of C_T that can be used to determine the value T that makes coordinates *independent* enough in a time delay reconstruction but still *correlated* to each other. In practice, one aims at optimizing the value of T suggested by both approaches. In any case, the embedding process is not so

sensitive to T, which makes it possible to use a T chosen by either approach where such T's are close.

Figure 2.3 illustrates the determination of the delay T using the autocorrelation and mutual information for two time series; a sinusoidal wave and white noise ($\mu = 0, \sigma = 1$)

2.1.2 Determining embedding dimension

(a) ***Singular Value Decomposition***

Singular value decomposition (SVD) attempts to find consistency in the results by trying a range of values of the embedding dimension d_e. One constructs a trajectory matrix from the data using a time delay $T = 1$. By decomposing the trajectory matrix into orthogonal coordinate space, one hopes that the corresponding distribution of singular values persist for all embedding dimensions greater than some minimum value. Thus, a rational basis for selecting the embedding dimension is established. However, Mees *et al.* (1987) have noted that the use of SVD in dimension estimation is limited because the number of singular values may depend on the details of the embedding and quality of the data as much as they do on the dynamics of the system.

(b) ***False Nearest Neighbours and False Strands***

Another method for determining the minimum embedding dimension is the false nearest neighbour (FNN) method of Kennel *et al.* (1992). The idea behind FNN is that if two points are neighbours in a reconstructed space of dimension d_e and fail to remain neighbours in dimension $d_e + 1$, then they are false. The necessary minimum embedding dimension occurs at that dimension when all true nearest neighbours are found, that is, they do not significantly separate in moving to a higher dimension. This assures that embedded points

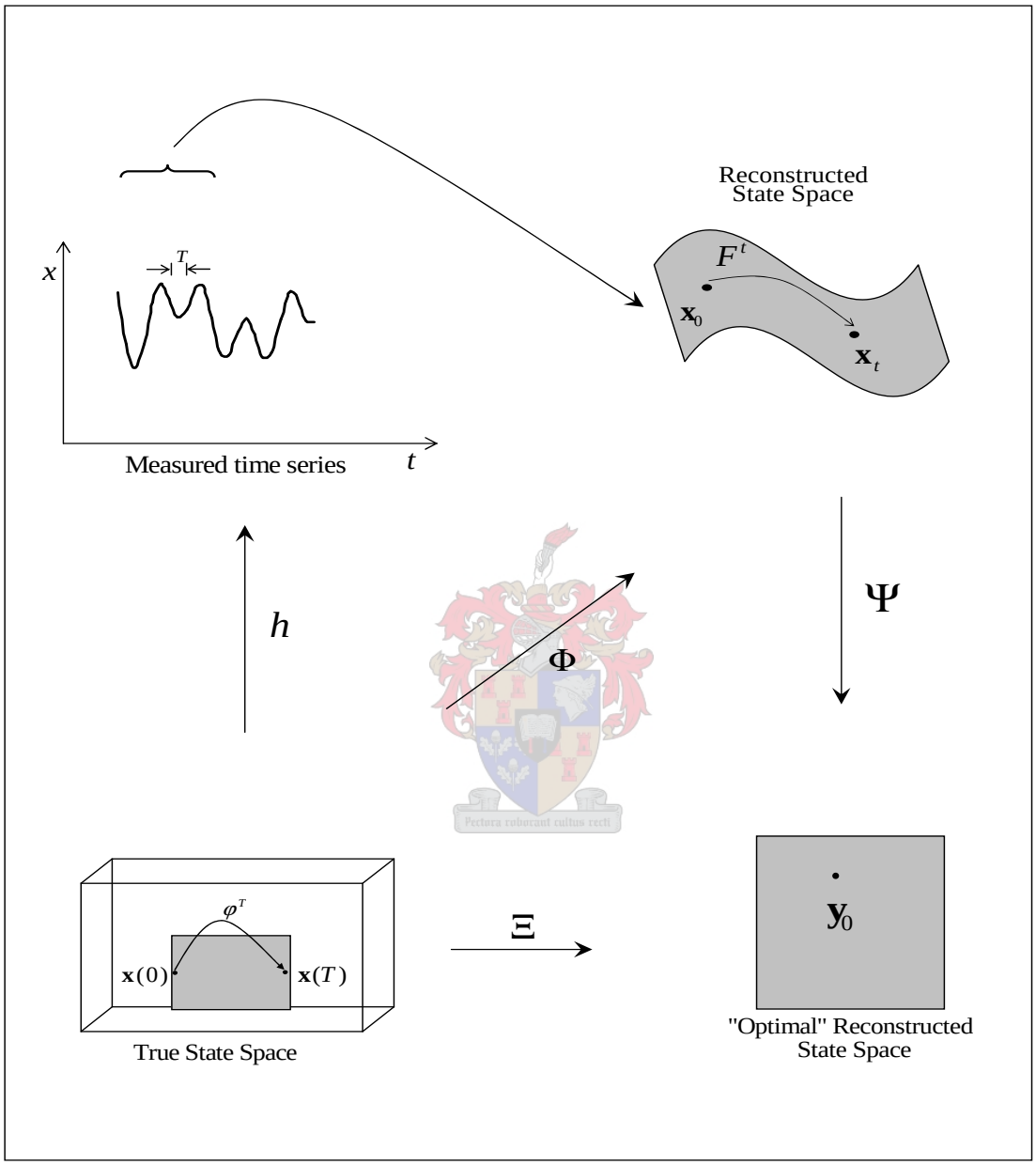


Figure 2.2: State space reconstruction. The underlying dynamical system φ, its states, $\mathbf{x}$, and the measurement function h are not observed. Measurements of the time series x separated by intervals of the lag time T form a delay vector $\mathbf{x} \in \mathbb{R}^m$ The delay reconstruction map Φ is defined $\Phi : \varphi \in \mathbb{R}^d \rightarrow F \in \mathbb{R}^m$. The coordinate transformation Ψ further maps the delay vector $\mathbf{x}$ into a new state space $y \in \mathbb{R}^{d'}, d' \leq m$. Adapted from Casdagli *et al.* (1991)

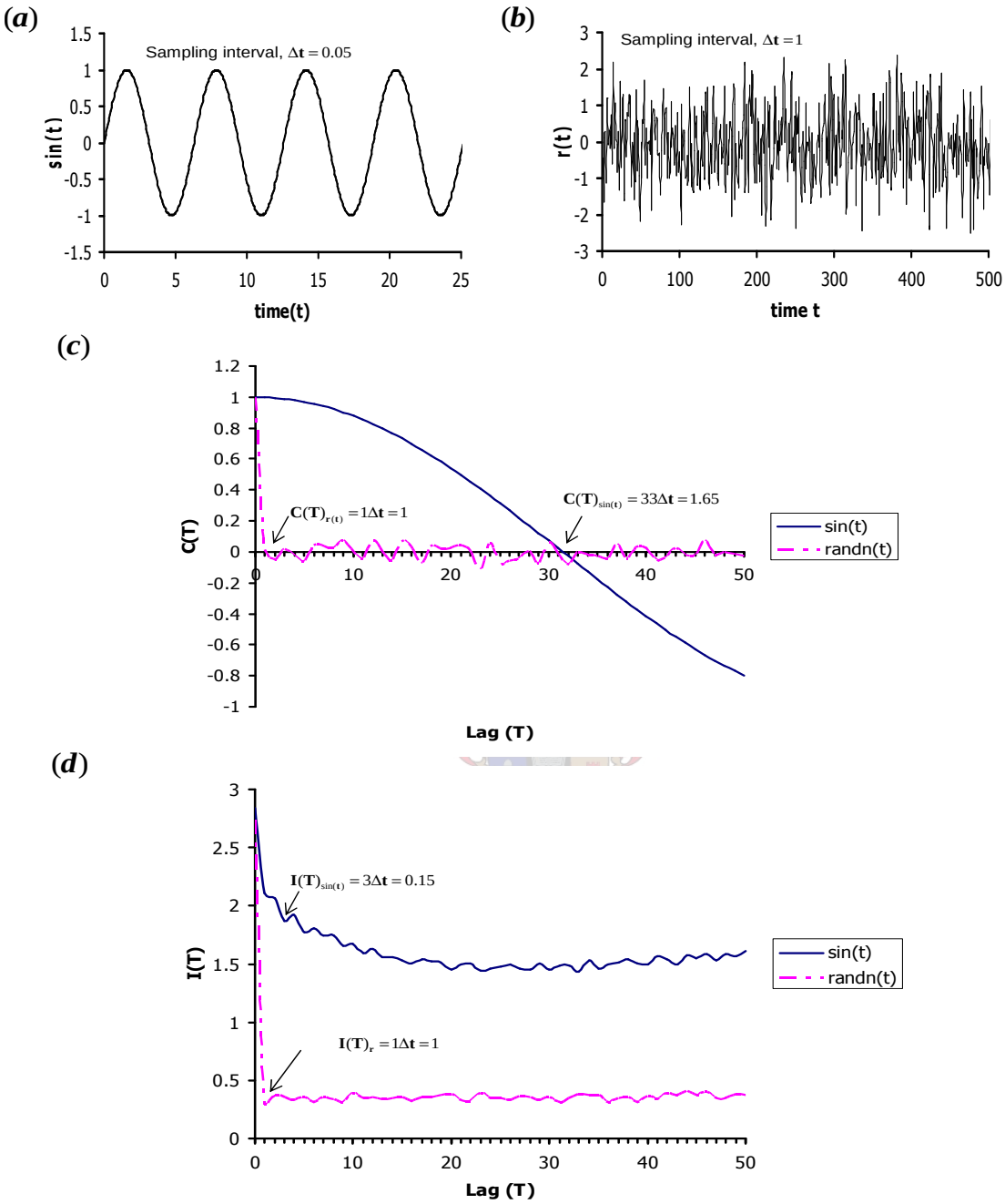


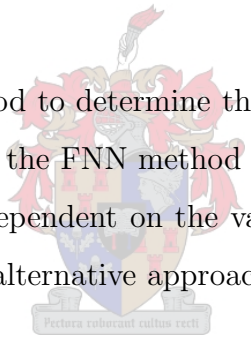
Figure 2.3: Illustration of determination of the delay T using autocorrelation and mutual information for a sine wave and Gaussian distributed variable. (a) Time series plot of the sine curve, with a sampling interval as indicated; (b) Time series plot of a normally distributed random variable; (c) Determining the delay lag T using autocorrelation function for the sine and random data; (d) Determining the delay lag T using mutual information criterion for the sine and random data.

have state space neighbours that are a result of the dynamics and not an artefact of being projected in low-dimensional space.

The method of false strands is an improvement on the FNN that eliminates various systematic effects that affect the FNN approach (Kennel and Abarbanel, 2002). The false strands method provides corrections to account for neighbourhood properties of oversampled data, the autocorrelation function for a small time delay, and sparsely populated regions of the attractor. These systematic effects make the determination of the necessary embedding dimension less certain when using the FNN method.

(c) ***Cao's Method***

Cao (1997) proposed a method to determine the minimum embedding based on the FNN method. Whilst the FNN method uses two pre-defined parameters, Cao's method is only dependent on the value of T. Also, it overcomes other shortcomings in other alternative approaches discussed above.



2.1.3 Embedding as a modelling problem

The formal approach to time series delay embedding using Takens' and related theorems may fail to provide an embedding useful for modelling, particularly when multiple timescales exist in the dynamics. Work done CADO[5] combines the embedding and modelling procedures into one procedure with a single optimization goal (Judd and Mees, 1995, 1998; Judd *et al.*, 1999; Small, 1998). This strategy is aimed at capturing the dynamics of the system in a model, which is a better measure for comparing models than just the prediction error.

The embedding dimension is implicitly found by determining the lag vector l_v.

[5]Center for Applied Dynamics and Optimization, University of Western Australia

This is achieved by constructing parameterized autoregressive models of the form

$$\mathbf{x}_t = \mathbf{AX} + \mathbf{e} \tag{2.9}$$

where

$$\mathbf{A} = [a_0, a_1, a_2, \ldots, a_n],$$

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x(n-1) & x(n) & \cdots & x(N-1) \\ x(n-2) & x(n-1) & \cdots & x(N-2) \\ \vdots & \vdots & \ddots & \vdots \\ x(0) & x(1) & \cdots & x(N-n) \end{bmatrix},$$

$$\mathbf{x}_t = [x(n), x(n+1), \ldots, x(N)],$$

$$\mathbf{e} = [x(n), x(n+1), \cdots, x(N)] - [\hat{x}(n), \hat{x}(n+1), \cdots, \hat{x}(N)],$$

where $\mathbf{e}$ is the residuals vector between actual and predicted values respectively. For zero-centered and normalized data, the constant a_0 and the corresponding rows of $1's$ in $\mathbf{X}$ fall off.

The autoregressive model in equation (2.9) has n parameters, of which some may only be necessary. By setting the parameters of those coefficients that contribute most in explaining the variation in the data to be non-zero and the rest to zero, the model can be re-formulated as

$$\hat{x}_t = a_0 + a_{l_1} x(t-l_1) + a_{l_2} x(t-l_2) + \cdots + a_{l_k} x(t-l_k) + e_t \tag{2.10}$$

for $t = n+1, n+2, \ldots, N$, where,

$$1 \leq l_1 < l_2 < \ldots < l_k \leq n.$$

Setting some of $a_i's$ in Equation (2.9) to zero requires re-estimation of the remaining parameters. This is achieved in a consistent and systematic manner by testing the significance of all parameters and determining the insignificant terms that are set to zero.

Rissanen's minimum description length (MDL) principle (Rissanen, 1999) is one particularly appropriate method that can be used in deciding the parameterization. An implementation employed by Judd and Mees (1995, 1998) and Small (1998) will be used to determine the appropriate lag vector l_v for non-uniform embedding of each individual component. Furthermore, in the case of multivariate embedding, the concept will be extended to the multivariate case where the lag vector of the component with the largest time window is used for all the other components before concatenation of the embedding space. All the other lag vectors are implicitly contained in the largest time window. Non-uniform embedding selects multiple time scales in the signal, avoiding redundancy in the embedding. Hence, unlike the straightforward implementation of Takens' theorem, the embedding space selected as such is effectively reduced and it may not be necessary to define an optimal embedding space.

2.2 Limitations of Scalar State Space Reconstruction

The embedding theorems guarantee that all possible reconstructions are equivalent and independent of the delay time or embedding dimension, provided d_e is at least twice the fractal dimension of the attractor. With real time series, one is confronted with many practical problems. The most obvious is the lack of *a priori* knowledge of the original state space. Since d is not known, it is not clear what value of the embedding dimension d_e should be used. Also, the values of observed time series

in the embedding theorems are arbitrarily precise, giving arbitrarily precise states. Thus, the specific value of the delay time T used is arbitrary and any reconstruction that meets the genericity conditions is as good as any other. But in practice real time series invariably contain noise, and so does simulated data series as number representation in computers is non-uniform and limited.

Casdagli *et al.* (1991) showed that there are principal limitations in state space reconstruction from real time series, which are invariably contaminated with both observational and dynamical (or multiplicative) noise. As explained in the previous section, state space reconstruction utilizes the flow of information from the state variables to the observed variables. Projecting a d-dimensional original state space onto a D-dimensional measurement ($D < d$) obscures certain information. State space reconstruction recovers some of this information. Noise amplification will occur if there is higher uncertainty in the reconstructed state than the observed time series, that is, the system appears less deterministic than it would if more information is provided. Casdagli *et al.* (1991) observe that noise amplification effects depend on the measured quantity; observation of one quantity may give more information than another. Finite time series length restrict the choice of an optimal delay time. This makes reconstruction of little or no practical value in some cases.

The introduction of an artificial folding of the reconstructed attractor is a huge defect of the delay embedding method using a scalar observable, something that is ignored in most applications (Hegger *et al.*, 1997). From equations (2.4) and (2.6), the discrete time evolution of dynamics in the delay space is given by

$$\mathbf{x}_{n+1} \quad = \quad F(\mathbf{x}_n) \tag{2.11}$$

With the exception of the first (most recent) component $x_{n+1} = F_1(\mathbf{x}_n)$, all the other components of $\mathbf{x}_{n+1}$ are simply copied from $\mathbf{x}_n$. The first component contains the entire information of the dynamics and, therefore, all the nonlinearity. This results

in strong folding effects in the direction of this component. Hence, it is necessary to have a sufficiently long enough noise free time series to allow for a spatial resolution on which the deterministic structure becomes visible. In a nonlinear modelling context, Hegger *et al.* (1997) remarked that the standard least squares minimization-based modelling techniques have an inherent systematic bias in the reconstructed dynamics due to two problems; (a) the *error-in-variables* problem arising from the standard least squares assumption that the independent delay vectors, $\mathbf{x}_n$, are noise free; (b) temporal correlations between successive delay vectors are not included; instead the maximum likelihood principle assumes statistical independence of all points in order to yield unbiased results. Judd and Small (2000) have proposed a canonical variate analysis approach that minimizes the iterated-prediction errors thus introduced.

Another major criticism of Takens' embedding theorem is probably that it requires the dynamics to be deterministic, autonomous, and stationary. It is relatively simple to show that times series exhibit *non-stationarity*[6]. Stationarity is more difficult to establish (Stark *et al.*, 1997). Schreiber (1999) gives a number of current trends being taken in resolving the non-stationarity problem. Small and Harrison (2000) generalize nonlinear modelling techniques to allow for the extraction of time dependent features from a non-stationary time series. It is not within the scope of this work to investigate stationarity. All signals will be assumed stationary, or tested for stationarity using existing tools.

Despite these limitations useful methods for the choice of the embedding dimension and time delay have been developed (Abarbanel, 1996; Abarbanel *et al.*, 1993; Kantz and Schreiber, 1997; Shreiber, 1998). Algorithms used in this work will be referred to in appropriate instances.

[6]Non-stationarity can be defined as time dependent changes in system dynamics.

2.3 Multivariate Nonlinear Time Series Analysis

The disproportionate attention that scalar time series have received from the time series research community partly stems from the fact that the question of inferring from a single measured time series the original dynamical system is not yet fully resolved. The simplicity and intuitiveness of scalar embedding is also another reason why people use scalar observables, even if other observables measured simultaneously are available. Multivariate time series introduce another complication of redundancy. It is always easier to deal with smaller set of time series derived from the original larger set. This is tackled using principal component analysis or its variants and other techniques in linear signal processing approaches. For nonlinear dynamics, the problem is considerably more difficult. Information theoretic-based approaches (Paluš, 1995; Prichard and Theiler, 1995; Schreiber, 2000) can be generalized to detect redundancies in multivariate time series. Use of multivariate time series has been by investigated different authors before in correspondingly different contexts, which will be reviewed in the following paragraphs.

Determining the embedding dimension d_e and the time delay T for state space reconstruction effectively defines a time window length $T_w = d_e T$ which allows all the underlying state variables to act sufficiently and be reflected in the embedded vector. Given a single time series we have no knowledge of the extent to which the chosen observation is related to all the underlying variables. It seems reasonable then to consider at least more than a single observation. Considering the illustration in Figure 2.4, separate consideration of either the x or y process variables gives different time window lengths because of the inherent different scales and variations. Large time windows reduce the number of reconstructed state vectors, posing a risk of a sparsely populated embedding space. This affects most signal processing methods for determining model parameters or system invariants. Including both time series in the reconstruction potentially allows the effect of all underlying state

variables to become apparent. Additionally, the time window length is reduced resulting in an increase in the number of points in the reconstructed state space, thereby facilitating better accuracy in later computations.

In the preceding arguments it has been implicitly assumed that the systems under investigations are autonomous. Hence, the measured observable has always been the output. Input-output systems are an interesting form of multivariate time series, where, in addition to system output(s), the input signal is also measured. Casdagli (1992) investigated the application of multivariate nonlinear prediction in the case of input–output systems. He argued that given an input sequence driving a dynamical system, then the output sequence, which taken on its own may be *stochastic*, can become deterministic and even non-chaotic when considered together with the input. A theoretical framework for the deterministic modelling of input–output given a scalar output time series and an input (also called forcing or driving system) time series was thus suggested. Hunter (1992) successfully applied the method in the analysis of an experimental time series of driven systems. Figure 2.5 depicts a schematic input-output system showing exogenous inputs that are associated with any real system.

Analogous to the autonomous system represented in equation (2.4), the response of a driven system based on delay embedding of response and input can be similarly formulated:

$$\mathbf{x}_t = f(x_{t-T}, x_{t-2T}, \ldots, x_{t-d_eT}, u_t, u_{t-T}, \ldots, u_{t-d_kT}) \tag{2.12}$$

The embedding now includes d_k delays of the input to the dynamic system and d_e delays of the response of the system. According Casdagli (1992), a generalized Takens' theorem extension assures that a diffeomorphism exists between expression (2.12) and the state evolution of a driven system of dimension d using a maximum of $(2d_e + 1)$ lags of both input and response. The model formulation in equation (2.12) allows application of time-series models to complex driven sys-

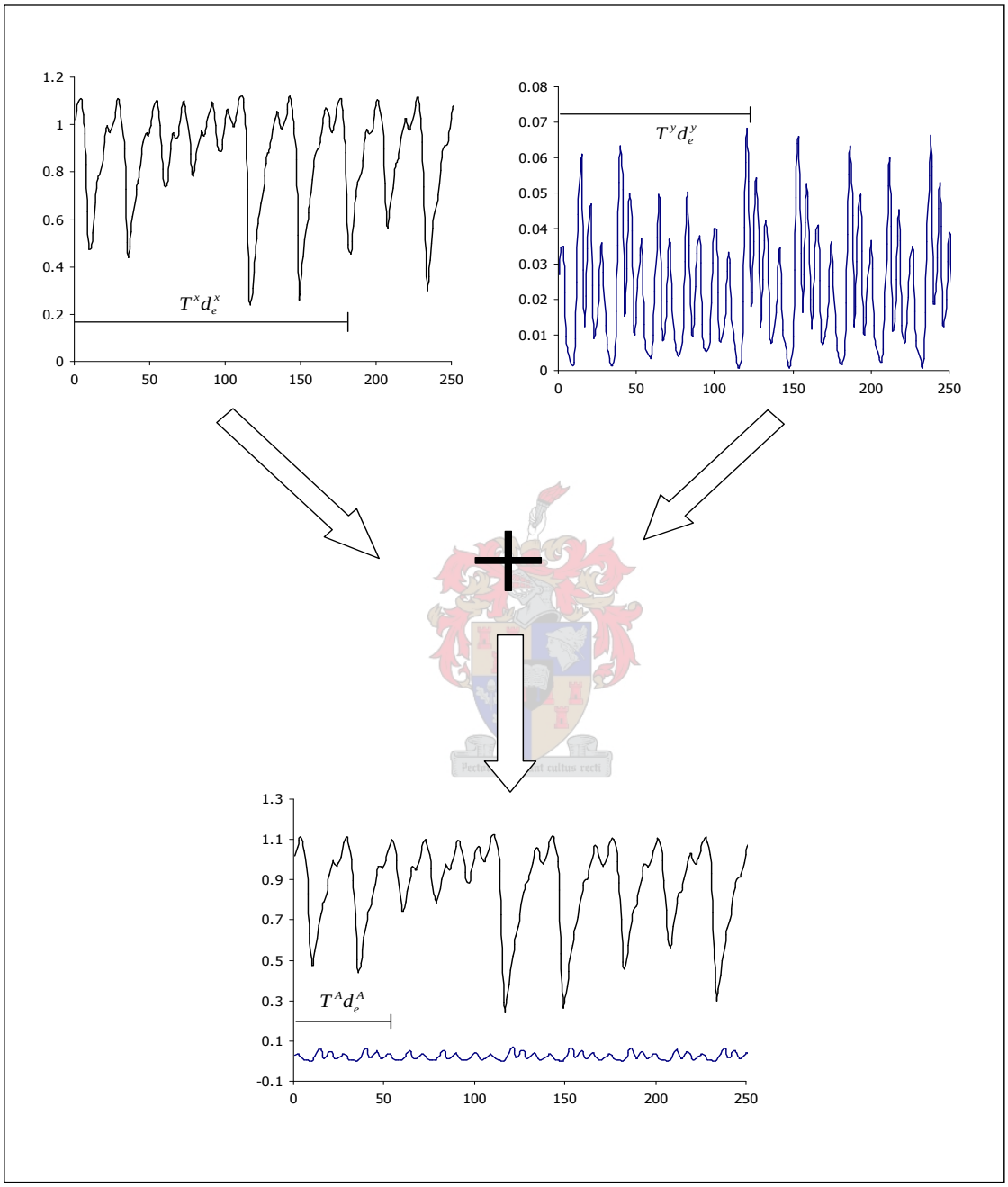


Figure 2.4: Reduction of time window length in multivariate embedding

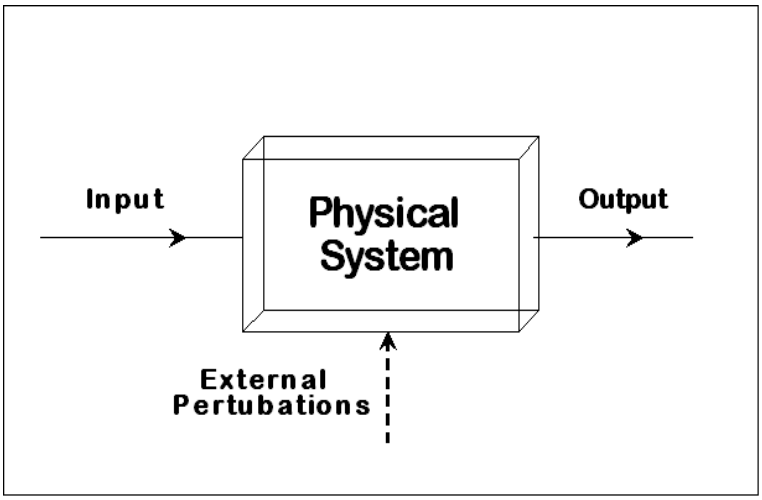


Figure 2.5: An input-output system

tems when the forcing signal is known. As pointed in Hunter (1992), knowledge of the input provides additional information that may simplify the modelling process, thereby improving the prediction process. Inclusion of the input signals in the reconstruction allows the model system behaviour to explore a wider region of space than is possible with only output signals.

Muldoon *et al.* (1998) proposed a new embedding theorem for the geometric analysis of time series perturbed by dynamical noise. Their theorem requires *so many* simultaneous measurements for an embedding of the original state space. It is not clear from their work how enough "so many" is and how one can determine this. Moreover, this qualification makes its practical application quite limited.

Porporato and Ridolfi (2001) developed a multivariate extension of a nonlinear time series-based method used in river flow forecasting. River flow forecasts are important in hydrology for the purposes of understanding river flow formation and prediction of flood events. Having previously shown that average daily river flows exhibited evidence of low-dimensional nonlinear determinism (Porporato and Ridolfi, 1996), they applied the method of nonlinear river flow forecasting using

temporal information from river upstream discharges, precipitation and temperature time series and spatial information from different points of a basin. It was shown that multivariate nonlinear prediction models performed better compared to univariate-based prediction models and provided flexibility to adapt to different sources of information, an important consideration in hydrology.

Cao *et al.* (1998) investigated the analysis of multivariate time series with a major focus on prediction, redundancy issues, the selection of the embedding dimensions, identification of functional relationships and synchronization between different variables. They proposed embedding each component of a multivariate time series using an optimal embedding obtained by minimizing the average prediction error of a nearest neighbour, locally constant predictor. Barnard (1999) and Barnard *et al.* (2001) argued that individual component embedding into state space risks significant statistical dependence amongst the reconstructed variables, resulting in a suboptimal reconstructed attractor. Instead, they proposed a method of multicomponent embedding that avoids both linear approximations in finding embedding dimensions and potentially suboptimal delay lags. After embedding individually each component, an optimal reconstructed state space was obtained using the independent components linear separation method. Their method displayed superior predictability than a simple scalar embedding approach. How and why their method reduces the risk of suboptimal delay lags is rather obscure since they determined the optimal lag for scalar reconstruction using the mutual information approach of Fraser and Swinney (1986), similar to the approach followed by Cao *et al.* (1998). An alternative approach proposed in this work is finding the optimal embedding window using the reduced autoregressive modelling approach used by Judd and Mees (1995, 1998) and Small (1998) in a univariate context.

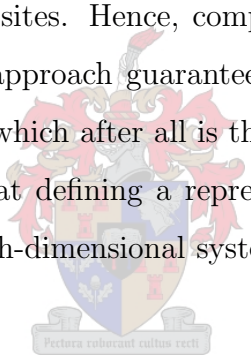
Hegger and Schreiber (1992) and Hegger *et al.* (1997) used multivariate time series within the context of noise reduction. They showed that multivariate data

when used in the noise reduction context are superior to scalar methods, in particular with respect to dynamical (multiplicative) error.

Wiesenfeldt *et al.* (1998, 2001) investigated the dependence between measured multivariate signals in the case of very weak coupling. They argued that where unidirectional coupling exists, a delay reconstruction using a time series from the response system with an embedding dimension that would be sufficient for the uncoupled or autonomous case captures all the dynamics of the driven system. To study the effect of including additional data in the reconstruction of a coupled system on predictability, *mixed states* containing delayed elements of both time series were used. They showed that in the case of strong coupling of two Hénon maps, both maps synchronized or exhibited same long-term behaviour of a property of their motion. Hence, it was concluded that use of the any of the variables in the reconstruction for specific embedding parameters is sufficient for predicting the one of the variables. Weakly coupled maps did not synchronize. In this case, the dynamics of one variable could not be determined uniquely by the dynamics of the other variable. It was shown that mixed states allow one to detect the existence and direction of weak coupling *below* the threshold of synchronization. This sensitivity makes their approach better than other earlier proposed measures of detecting weak coupling in physical systems.

There is growing interest in spatiotemporal series analysis, spurred largely by the introduction of the coupled map lattice (CML) as an alternative model for such systems by Kaneko (1989a,b). An extensive review on the state and progress in the study of spatiotemporal (ST) systems can be found in Cross and Hohenburg (1993). A spatiotemporal series consists of time series taken from different sites of a spatially extended system. Ørstavik *et al.* (2000, 1998) have studied multivariate time series generated by coupled map lattices showing spatiotemporal chaos for the purposes of reconstruction, cross-prediction and estimation of intensive quantities. Unlike

purely temporal or purely spatial time series, ST series are not low-dimensional. The coupling effect induced by the infinite number of interacting state variables arising from the infinite spatial extent results in a high-dimensional system. Characterization of such systems poses great challenges. Beyond mere academic pursuit, spatiotemporal systems are more prevalent in nature than purely temporal systems. It can be argued that chemical, metallurgical, and most physical systems have a spatial extent in their configuration. For example, in the flotation process used in mineral beneficiation, the performance of cleaner cells is affected by inflow of concentrate from rougher cells, whose performance in turn is affected by grinding mill overflow and scavenger returns. Furthermore, within each cell various microscopic reactions occur in different sites. Hence, complete analysis requires taking into account all these effects, an approach guaranteed to be self-defeating at the onset! Simplification is necessary, which after all is the whole point of engineering. Spatiotemporal analysis is aimed at defining a representation that allows for the tractability of these ill-defined, high-dimensional systems.



2.4 Nonlinear Statistics of Dynamical Systems

A broadband power spectrum is indicative of an underlying infinite dimensional system (noise) or a system that evolves nonlinearly on a finite-dimensional attractor. Both alternatives are equally valid. Therefore, there is a need to derive dynamical statistical quantities that distinguish between noise and low-dimensional determinism. Derivation of such quantities is based on the *ergodic theory* of nonlinear dynamical systems (Eckmann and Ruelle, 1985; Kantz and Schreiber, 1997). This section discusses the theory and concepts that used in characterizing dynamical systems.

An attractor $\mathcal{A}$ is a global description of the asymptotic or long-term behaviour

of a dynamical system. A detailed description at smaller scales is given by a probability measure ρ on $\mathcal{A}$, which describes the frequency with which various parts of $\mathcal{A}$ are visited by the orbit $t \rightarrow \mathbf{x}$ describing the system. ρ is conveniently defined as a time average at points $\mathbf{x}$ in phase space and is invariant under the action of the dynamical system. A family of invariant quantities exist that are useful in the characterization of deterministic systems. These quantities share the property of *invariance under smooth transformation of the state space.* An ergodic property is invariant under smooth coordinate transformations and assumes the same value for almost any set of initial conditions with respect to the Lebesgue measure. An invariant probability measure on the attractor may be decomposable into several different invariant pieces (Eckmann and Ruelle, 1985). An *indecomposable* invariant probability measure is said to *ergodic.* An irreducible natural measure can thus be defined on an attractor

Definition 2.1 *Let $\rho(d\boldsymbol{x})$ be the average time a typical trajectory spends in the phase space element $d\boldsymbol{x}$ of some continuous function φ. If ρ is ergodic then the ergodic theorem asserts that for every continuous function φ,*

$$\begin{aligned} \rho &\equiv \int \rho(d\boldsymbol{x})\varphi(\boldsymbol{x}) \\ &= \lim_{N\rightarrow\infty} \frac{1}{N} \int_0^N \varphi[\boldsymbol{x}(t)]dt \end{aligned} \tag{2.13}$$

for almost initial conditions with respect to the measure ρ; i.e., a space average equals a time average.

Characterization of nonlinear dynamical behaviour proceeds by calculating the values of quantities that are insensitive to changes in the initial conditions on the trajectory, or when perturbations are encountered. These quantities include the spectrum of Lyapunov quantities that characterize the stability of trajectories; di-

mensions that describe topological properties defined on $\mathcal{A}$; and entropies, relating the information properties of the invariant measure on the attractor.

2.4.1 Dimensions

Dissipative systems have asymptotic behaviour defined on a finite-dimensional, invariant subset of the corresponding state spaces. Dimensions are a numeric description of these invariant subsets. The phase space dimension is the number coordinate points needed to completely specify the state of the system at any instant. For systems defined by ordinary differential equations or discrete mappings, the dimension corresponds to the number of equations defining the evolution of each state variable. Systems defined by partial differential equations have an infinite phase space dimension. However, as already mentioned above, the asymptotic behaviour of dissipative dynamical systems relaxes onto a low-dimensional, invariant subset of the complete state space. It is found that certain nonlinear systems have a non-integer dimension due to the fractal nature of the attractor. Various metrics have been defined for estimating this non-integer dimension, of which the important ones are discussed next.

(a) ***The box-counting dimension***

Hausdorff (Eckmann and Ruelle, 1985) gave the first definition of a non-integer dimension that has formed the basis for other definitions. The *Hausdorff dimension* characterizes the self-similarity of sets. However, due to inherent practical computational limitations, the closely related *capacity or box-counting* dimension is often used. It gives an upper bound on the Hausdorff dimension. The capacity dimension is based on the scaling properties of the attractor with size. Given a point set in $\mathbb{R}^m$, one superimposes hypercubes or boxes with side length ε on the space in which the set resides. For a *self-similar* set, the number of boxes $M(\varepsilon)$ that contain at least a single point

scales as,

$$M(\varepsilon) \quad \propto \quad \varepsilon^{-D_F} \tag{2.14}$$

The *capacity dimension* D_F can be shown to be defined by

$$D_F \quad = \quad \lim_{\varepsilon \to 0} \frac{\log_e(1/M(\varepsilon))}{\log_e \varepsilon} \tag{2.15}$$

D_F is a metric of the dimension of the space in which the set lives. Generally, the box-counting dimension is computed for increasing embedding dimensions. For embedding dimensions less than or equal to D_F, the attractor's projection "fills" the embedding space, giving an estimated fractal dimension equal to the embedding dimension. As the embedding dimension increases through the minimum required for complete unfolding of the geometric structure, the calculated fractal dimension saturates. The estimate for D_F is taken as the minimum value of d_e at which the saturation is first observed.

(b) ***Generalized dimensions***

The capacity dimension considers only the geometrical structure of the attractor, ignoring the distribution of points on the attractor. It is logical to give more weight, and thus larger proportions of the natural measure, to the most frequently visited regions (Kantz and Schreiber, 1997). A certain family of dimensions called the *generalized or Renyi* dimensions differ in the way weights are assigned to regions with different densities. The generalized correlation integral is defined as

$$C_q(\varepsilon) \quad = \quad \int_{\mathbf{x}} p(\mathbf{x}_\varepsilon^{q-1}) d\rho(\mathbf{x}) \tag{2.16}$$

For a self-similar point set C_q scales with ε as

$$C_q(\varepsilon) \quad \propto \quad \varepsilon^{(q-1)D_q}, \qquad \varepsilon \to 0 \tag{2.17}$$

The generalized dimensions D_q are then determined according to

$$D_q = \lim_{\varepsilon \to 0} \frac{1}{q-1} \frac{\log_e C_q(\varepsilon)}{\log_e \varepsilon} \tag{2.18}$$

The generalized dimension (D_0) is equivalent to the box-counting dimension D_F. The information dimension (D_1) quantifies the average amount of information needed to specify a state at some specified resolution ε

$$D_1 = \lim_{\varepsilon \to 0} \frac{\sum_i p_i \log p_i}{\log \varepsilon} \tag{2.19}$$

where p_i is the probability of a point being in the ith partition (Abarbanel, 1996; Eckmann and Ruelle, 1985; Kantz and Schreiber, 1997). D_1 gives information on the growth of the information required as ε decreases.

The correlation dimension D_2 is defined by

$$D_2 = \lim_{\varepsilon \to 0} \frac{-\log \sum_i p_i^2}{\log \varepsilon} \tag{2.20}$$

The numerator constitutes a two-point correlation function which measures the probability of finding a pair of randomly chosen points within a given partition element. The Grassberger–Procaccia (GP) estimate to D_2 is found by estimating the term $\sum_i p_i^2$ using the correlation sum (Grassberger and Procaccia, 1983).

$$C(\varepsilon) = \frac{1}{N(N-1)} \sum_{i \neq j} \Theta(\varepsilon - \| \mathbf{x}_j - \mathbf{x}_i \| \tag{2.21}$$

To estimate the dimension, a series of $\log C(\varepsilon)$ versus $\log \varepsilon$ curves for increasing embedding dimensions are plotted. At higher embedding dimensions the slope of the $\log C(\varepsilon)$ versus ε-plot saturates (for dissipative low-dimensional systems) at a value equal to the attractor's dimension. Although the dimension is defined as the slope of this plot in the limit $\varepsilon \to \infty$, this region is dominated by noise and the effects of discrete measurement channels. Hence,

it is hoped to identify a scaling region within some intermediate length scales, where a constant slope allows reliable estimation of the dimension. (See also a section on estimating correlation dimension in appendices.)

Amongst the generalized dimensions D_2 is the easiest to determine numerically and, therefore, is used extensively. However, in general $D_2 \leq D_1 \leq D_0$ holds for the three dimension measures. The equality conditions apply when the points are distributed uniformly over the attractor (Grassberger and Procaccia, 1983).

Although the GP method gives a reasonable estimate for D_2 it has been pointed out that the information it gives is limited because of the following assumptions (Judd, 1992);

(i) the calculation of the interpoint distances equation (2.21) assumes independence of the interpoint distances;

(ii) the correlation sum in equation (2.21) is a smooth function without statistically correlated information, and;

(iii) all the information about dimension is contained within the scaling region to which a straight line is fitted. However, the scaling region may only reflect large-scale properties of the attractor and not the underlying dimension.

Another major failing of the GP method is that it does not provide error bars on the dimension estimate. Judd (1992) proposed an estimator based on the Grassberger-Procaccia idea of calculating interpoint distances from a trajectory that considers the distribution of the interpoint distribution directly instead of the correlation sum. However, the method still suffers from the problem of the dependence of the interpoint distances. In spite of this, Judd's

method gives a reliable estimator for the correlation dimension, particularly for attractors with $D_2 < 4$.

Higher order generalized dimensions with $q \geq 3$ are also defined but will not be discussed here as there are not practically useful due to the computational difficulties they present.

2.4.2 Lyapunov exponents

The box-counting dimension gives information on the support in which the attractor lives whilst the generalized dimensions D_q characterize the asymptotic spatial distribution of points along a trajectory. However, both do not provide information on the dynamic, temporally evolving structure of the system, which is more useful for practical applications such as modelling and control. Lyapunov or characteristic exponents describe the time-ordered points on a trajectory and, thus, the dynamics defining the evolution of the trajectories. A typical characteristic of chaotic dynamical systems is their sensitivity to initial conditions, that is, two initially infinitesimally close points in state space diverge exponentially with time.

In dissipative systems a folding mechanism constrains the trajectories to a low-dimensional attractor. The two effects of exponential separation of initially similar points and the simultaneous constraining of the region of space are sometimes referred to as the *stretching and folding* mechanism. Thus, whilst the attractor's overall volume contracts, expansion occurs in certain directions of phase space. The total rate of contraction in the other directions is obviously much greater than the rate of expansion.

There are as many Lyapunov exponents as there are phase space directions. The spectrum of Lyapunov exponents is determined by following the time evolution of initially similar points $\mathbf{x}_n$ and $\mathbf{x}'_n$ in the tangent space separated by an infinitesimal distance δ_n. The time evolution of their separation distance under the action of the

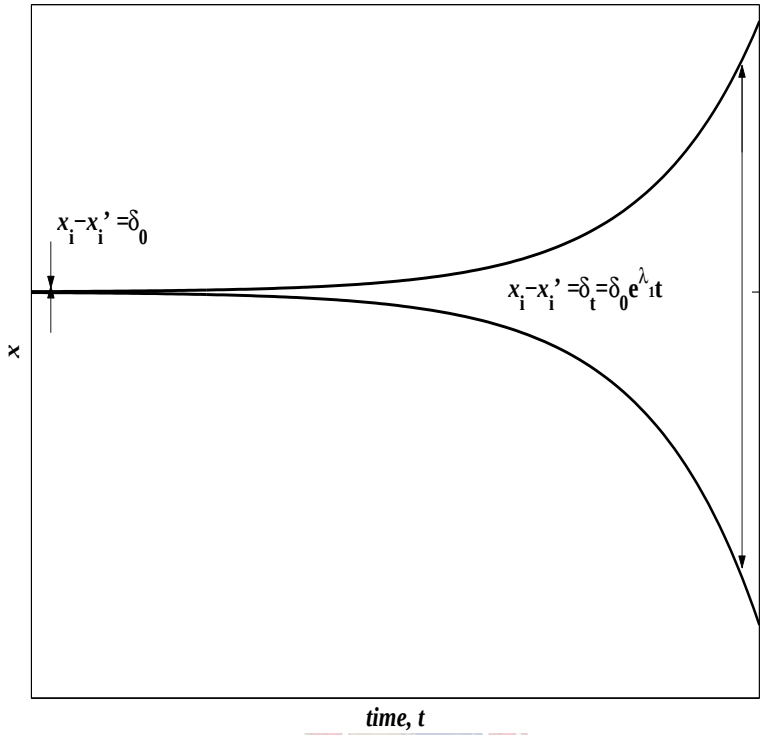


Figure 2.6: Exponential divergence of initially infinitesimally close points

dynamical system $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is given by the a Taylor series expansion of $f(\mathbf{x}'_n)$ around $\mathbf{x}_n$

$$\begin{aligned} \mathbf{x}_{n+1} - \mathbf{x}'_{n+1} &= f(\mathbf{x}_n) - f(\mathbf{x}'_n) \\ &= \mathbf{J}_n(\mathbf{x}_n - \mathbf{x}'_n + O(\| \mathbf{x}_n - \mathbf{x}'_n \|^2) \end{aligned} \tag{2.22}$$

where $\mathbf{J}_n = \mathbf{J}(\mathbf{x}_n)$ is the Jacobian square matrix that gives the linearized dynamics of f centered at $\mathbf{x}$. Given $\delta_n = \mathbf{x}_{n+1} - \mathbf{x}'_{n+1}$ it is possible to compute its modulus a time step later. Letting $\mathbf{e}_i$ and Λ_i be the eigenvectors and eigenvalues respectively of $\mathbf{J}$, eigenvector decomposition of $\mathbf{J}$ gives

$$\delta_{n+1} = \sum_i c_i \Lambda_i \mathbf{e}_i \tag{2.23}$$

where c_i are coefficients and Λ_i are local stretching factors. Each arbitrary point of the phase space has different eigenvectors and eigenvalues. Global eigenvalues can be defined by taking a proper average over the different local stretching factors. The *Lyapunov exponent* λ_i is the normalized logarithm of the modulus of the ith eigenvalue Λ_i of the product of all Jacobians along the trajectory (in time order) in the limit of an infinitely long trajectory;

$$\lambda_i = \lim_{N\to\infty} \frac{1}{N} \log_e |\Lambda_i^{(N)}| \tag{2.24}$$

where Λ_i is given by

$$\prod_{n=1}^{N} \mathbf{J}_n \mathbf{u}_i^{(N)} = \Lambda_i^{(N)} \mathbf{u}_i^{(N)} \tag{2.25}$$

An m-dimensional space has m Lyapunov exponents corresponding to each of the coordinates. The Lyapunov exponents are typically arranged in a decreasing order; $\lambda_1 > \lambda_2 \geq \ldots > \lambda_m$. A necessary condition for the existence of a chaotic attractor requires that $\lambda_1 > 0$. If $\lambda_i < 0$ for all i then the attractor is a stable fixed point and has dimension zero, whilst if the non-negative exponents are zero, the attractor is a limit cycle. Multiple null exponents correspond to the number of incommensurate frequencies in a quasi-periodic system, which is also the system's dimension. The set of the all the exponents is called a *Lyapunov spectrum*. λ_1 measures the rate of divergence of two points initially separated by an infinitesimal distance δ_0. For a chaotic system, the separation distance increases as $\delta_t = \delta_0 e^{\lambda_1 t}$ after time t as shown in Figure 2.6.

Derivation of the whole Lyapunov spectrum enables the computation of other system invariants. For example, the Kaplan-Yorke conjecture relates the dimension of the attractor to the stability properties of the system dynamics and is given by (Abarbanel, 1996; Eckmann and Ruelle, 1985):

$$D_L = K - \frac{\sum_{i=1}^{K} \lambda_i}{\lambda_{K+1}} \tag{2.26}$$

such that $\sum_i^K \lambda_i > 0$, and $\sum_{i=1}^{K+1} < 0$. D_L is preferable when estimating high dimensions. Pesin's identity relates the sum of exponents to the Kolmogorov-Sinai entropy (described in subsection 2.4.3):

$$h(X) = \sum_{\lambda_i > 0} \lambda_i \tag{2.27}$$

Computation of the entire Lyapunov spectrum space is difficult when all we have is a measured time series. Reconstruction of the state space results in spurious exponents since $d_e > d$. Typically, only λ_1 is computed for positive identification of chaos in physical system.

2.4.3 Entropy

A system with sensitive dependence on initial conditions generates information since two initially infinitesimal close points indistinguishable at a given resolution evolve into distinguishable states after a finite time. The concept of entropy characterizes the information flow from the state of the system to the observations (Kantz and Schreiber, 1997). Assuming observations on a system follow some probability distributions, transitions between different states occur with well-defined probabilities. Hence, it is worthwhile knowing how much information, on average, does a single measurement provide about the state of the system or how much information can be deduced about the future observations given past observations.

For static distributions, the order-q Renyi entropies characterize the amount of information needed to specify the future value an observable with a certain precision if we know the probability density function that a certain value is attained. The concept of mutual information discussed earlier is a special case of these entropies. In general, the mutual information, or *redundancy*, can be defined for any m variables and their corresponding distributions.

By including correlations in time or transition probabilities, we can define the

Kolgomorov-Sinai (KS) invariant $h(\rho)$, which measures the average rate of information generation, where ρ is an ergodic probability measure for a dynamical system. The KS entropy is important in the characterization of dynamical systems because; (a) it gives information on the precision required for the predictability of the system, and ; (b) it supplies topological information on the folding process, that is, the flow of information between the small and large scales. As useful as they are, it is numerically difficult to calculate entropy values from time series data (Abarbanel, 1996; Eckmann and Ruelle, 1985; Kantz and Schreiber, 1997). Nevertheless, where estimates can be obtained the following generalizations are observed – linear deterministic systems typically have a zero-entropy value whilst linear stochastic systems have an infinite entropy. Finite entropy estimates are obtained for nonlinear deterministic systems.

2.5 Dimensionality Reduction

Developments in online sensor technology have facilitated the simultaneous measurement of different process variables to aid in system monitoring, control, and detection of abnormal behaviour. Large historical data repositories now exist for the purpose of data mining and retrieval of critical knowledge of process behaviour. Given the low signal-to-noise ratios associated with measuring only a single variable it is advantageous to perform multivariate measurements. However, multiple signals are highly correlated and noncausal in nature. Collinearity introduces complications in signal processing, of which the *"curse of dimensionality"*[7] phenomenon is an example. Furthermore, high-dimensional spaces are inherently *sparse* and, coupled with linear correlations, results in a large mean sum square error for probability

[7]That is, the sample size needed for the parameterization of a function of several variables (to a specified tolerance) increases exponentially with the number of variables.

density and function estimation (Eneva *et al.*, 2002).

The apparent high-dimensionality of a process is usually due to the influence of exogenous inputs and irrelevant process variables. The underlying dynamical processes of many systems are governed by a very small number of degrees of freedom (typically less than 5). The challenge then is how to exploit the inherent redundancy and reduce the effect of noise in the observed variables. It turns out that the true state coordinates, called *hidden* or *latent variables*, are reflected in the measured process variables according to some *unknown* combination or mixing rule. As discussed previously (section 2.1), reconstruction of state space variables uses dynamical information in time-lagged values of a signal resulting in an embedding whose dimension is more than twice the intrinsic or fractal dimension. The reconstruction procedure also gives rise to an additional symmetry. For example, for a 3-dimensional attractor, only two of the three projections onto the planes of the reconstructed space are different[8] (Landa and Rosenblum, 1991). Hence, correlations existing between variables in both multivariate measurements and reconstructed multidimensional state spaces must be exploited appropriately for data mining.

It is possible to circumvent the problems associated with high-dimensional data by determining a low-dimensional space that adequately describes the variations in the original variables. A popular approach is *dimensionality reduction* (Eneva *et al.*, 2002). Dimensionality reduction seeks an encoding function $\mathcal{G}$ and a decoding function $\mathcal{F}$ such that the following mappings hold;

$$\mathcal{G}(\mathbf{x}) : \mathbb{R}^m \rightarrow \mathbb{R}^k \tag{2.28}$$

$$\mathcal{F}(\mathbf{z}) : \mathbb{R}^k \rightarrow \mathbb{R}^m \tag{2.29}$$

where $k \ll m$. Thus, for every input vector $\mathbf{x} \in \mathbb{R}^m$, a low-dimensional equivalent

[8]For comparison, a trigonometry perspective, the triangle inequality states that the distance between points is less than or equal to the sum of distances from a third point.

transformation is obtained by $\mathbf{z} = \mathcal{G}(\mathbf{x}) \in \mathbb{R}^k$. The inverse transformation $\mathcal{F}$ maps the data back to the original input space. Latent variables projection methods are commonly used to in dimensionality reduction. These methods exploit the redundancy introduced through correlations in defining an effective dimensionality which is less than original data dimensionality. Two such latent variables projection methods are (linear) principal component analysis (PCA) and independent component analysis (ICA).

(a) ***Principal Component Analysis (PCA)***

PCA transforms a data set of n observations of m-dimensional correlated variables $\mathbf{X}$ into *uncorrelated* $k-$dimensional latent variable scores $\mathbf{T}$;

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2.30}$$

where $\mathbf{X} \in \mathbb{R}^{n\times m}, \mathbf{T} \in \mathbb{R}^{n\times k}, \mathbf{P} \in \mathbb{R}^{m\times k}$. $\mathbf{P}$ is called the *loading matrix* that shows how the latent variables are related to the original variables, and $\mathbf{E}$ is a matrix of residuals not explained by the PCA model (Jollife, 1986). The transformation in equation (2.30) assumes that each of the original variables has a Gaussian distribution. Even in the case of non-Gaussian variables the central limit theorem from probability asserts that the sum of independent variables tends towards a Gaussian distribution. Hence, in principle it is still possible to apply the transformation to variables with a non-normal distribution.

PCA finds linear combinations of the variables via an eigenvector decomposition of the covariance matrix obtained from the measured variables. The kth principal component is defined as the linear combination $\mathbf{t}_k = \mathbf{X}\mathbf{p}_k$ that has maximum variance subject to $|\mathbf{p}_k| = 1$ (without loss of generality, we assume $\mathbf{X}$ is zero-centered and normalized). The principal component loading vectors $\mathbf{p}_k$ are the eigenvectors of the sample space $\mathbf{X}$ with a covariance matrix Σ whose spectral decomposition is $\Sigma = \mathbf{P}\Lambda\mathbf{P}^T$, where $\mathbf{P}$ is orthonormal and

Λ diagonal. The elements of Λ are arranged in decreasing order and satisfy $\lambda_k = \mathrm{var}(\mathbf{t}_k)$.

Most of the linear variability in the data is captured in the first few principal components, hence one needs to compute only k significant eigenvectors.The significant principal components can be evaluated using, for example, cross-validation methods. The description of the data is then given by a decomposition into the sum of an outer product of vectors $\mathbf{t}_i$ and $\mathbf{p}_i$ and a residual matrix $\mathbf{E}$,

$$\mathbf{X} = \sum_{i}^{k} \mathbf{t}_i \mathbf{p}_i^T + \mathbf{E} \tag{2.31}$$

Such a linear mapping has the least sum of squared errors and the maximum mutual information between the original vectors and their projections $I(x, x*) = \frac{1}{2} \ln\left(\prod_{i=1}^{k} \lambda_i\right)$.

The transformation in equation (2.30) assumes that each of the original variables has a Gaussian distribution. However, even in the case of non-Gaussian variables the central limit theorem from probability asserts that the sum of independent variables tends towards a Gaussian distribution. Hence, in principle it is still possible to apply the transformation to variables with a non-normal distribution.

(b) ***Independent Component Analysis (ICA)***

ICA is a statistical method for transforming a multidimensional random vector into components that are statistically *independent* from each other (unlike PCA which seeks a transformation that gives only *uncorrelated* components). The observed process variables $\{x_i\}_{i=1}^{n}$ are assumed to be n linear or nonlinear mixtures of n *independent* components z_i generated by a mixing model defined

by

$$\begin{aligned} x_i &= a_{i1}z_1 + a_{i2}z_2 + \ldots + a_{in}z_n, \qquad \forall i \\ \text{or} \quad \mathbf{x} &= \mathbf{A}\mathbf{z} \in (\mathbb{R}^{n\times n} \circ \mathbb{R}^{n\times 1}) \end{aligned} \tag{2.32}$$

where x_i and z_k are random variables (Hyvärinen and Oja, 1997, 2000). Under the assumption of independence, the mixing matrix $\mathbf{A}$ is estimated and its inverse, the separation matrix $\mathbf{W}$, obtained. The independent components are then found as

$$\mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{E} \tag{2.33}$$

As in the PCA case, the separation matrix $\mathbf{W}$ cannot be found exactly because we have no knowledge of $\mathbf{A}$. ICA finds an estimate that gives a good approximation to $\mathbf{W}$ by maximizing the *nongaussianity* of $\mathbf{W}\mathbf{x}$ – "nongaussianity is independence". The independent components *must* be non-Gaussian for ICA to apply. Thus, a measure of non-Gaussianity has to be formulated. Kurtosis and negentropy are commonly used measures of the deviation of the distribution of data from the bell-shaped normal curve.

Kurtosis measures the extent to which values cluster around a central point. It is given by the fourth-order moment function,

$$\alpha_4 = E\{x^4\} - 3(E\{x^2\})^2 \tag{2.34}$$

For Gaussian random variables,$\alpha_4 = 0$, otherwise it is positive or negative. However, kurtosis is very sensitive to outliers and therefore not a robust measure of non-Gaussianity.

The entropy of a random variable gives the amount of information a variable possesses and is defined according to;

$$H(\mathbf{X}) = -\sum_i p(x_i)\log p(x_i) \tag{2.35}$$

where $\mathbf{X}$ is a discrete random variable and $p(x_i)$ the probability that the random variable $\mathbf{X}$ assumes the value x_i. For continuous random variables the summation in equation (2.35) is replaced by the integral operator.

Random variables have the largest entropy among all random variable of equal variance because of their unpredictability and lack of structure. Entropy can be used as a measure of non-Gaussianity by setting the entropy of a random variable to be equal to zero. Thus, all other random variables have a negative entropy

$$J(\mathbf{x}) \quad = \quad H(\mathbf{x}_{gauss}) - H(\mathbf{x}) \tag{2.36}$$

where $\mathbf{x}_{gauss}$ is a Gaussian random variable with the same covariance matrix as $\mathbf{x}$. It can be shown that, for normalized and uncorrelated x_i, the mutual information of n random variables $\{x_i\}_{i=1}^{n}$ are related to negentropy as

$$I(x_1, x_2, \ldots, x_n) \quad = \quad C - \sum_i J(x_i) \tag{2.37}$$

where C is constant independent of the separation matrix $\mathbf{W}$.

The independent components $\mathbf{z}$ are obtained from the invertible transformation $\mathbf{z} = \mathbf{W}\mathbf{x}$, where $\mathbf{W}$ is determined such that the mutual information of the transformation components is minimized (under the decorrelation constraint). In the *fixed-point* algorithm (Hyvärinen and Oja, 1997, 2000), this is achieved by finding the coordinates in which the negentropy is maximized;

$$\begin{aligned} & \max \textstyle\sum_i^M J_G(\mathbf{w}_i) && \text{wrt. } \mathbf{w}_i = 1, \ldots, n \\ \text{subject to} \quad & E\{(\mathbf{w}_k^T\mathbf{x})(\mathbf{w}_j^T\mathbf{x})\} = \delta_{jk} \end{aligned} \tag{2.38}$$

where,

$$J_G(\mathbf{w}) \quad = \quad k[E\{G(\mathbf{w}^T\mathbf{x}\} - E\{G(\nu)\}]^2, \tag{2.39}$$

G is some smooth, symmetric *contrast* function that estimates the probability density function of an independent component, k an insignificant positive constant, and ν a standardized Gaussian variable.

Thus, using ICA an optimally reduced optimal separated basis space can be found. The implementation of ICA requires *sphering* or *whitening* of the multidimensional variable[9]. Further details on independent component analysis can be found in Hyvärinen and Oja (1997, 2000).

Invariably, process variables are non-Gaussian because of the nature and structure of operating conditions. Hence, the application of ICA is more or less always successful. However, there has been little study done on the effect of the separation approach used, i.e. PCA and ICA, on the resulting application of the reduced space. Part of this study will investigate the effect the separation approach has on the resultant predictive models designed for control purposes.

2.6 Concluding Remarks

Low-dimensional determinism offers an alternative description for observed irregular behaviour in physical systems. Using the embedding theorems and a scalar time series taken on the system it is possible to reconstruct a dynamical system's attractor equivalent to the true underlying attractor. Characterization of such systems involves use of nonlinear dynamical invariant quantities defined on the attractor. Three commonly used quantities are the dimension estimates, Lyapunov exponents, and entropy, which quantify topological, geometrical, and information properties associated with a dynamical system respectively. However, finite and imprecise data invariably available introduce principal limitations in the applications of the embedding approaches. Furthermore, use of a scalar observable introduces artificial

[9]Sphering is scaling of the all eigenvalues of a matrix to unity

folding in the reconstructed attractor. Although the problem of inferring system dynamics is not yet fully well-established, especially in industrial applications, the prevalence of multivariate measurements on process systems motivates the exploitation of these repositories of data for improved process understanding and operation. Multivariate analysis introduces its own complications of collinearity, synchronization, etc., that need to be resolved. These issues have been treated extensively especially in linear approaches using, for example, principal component analysis. Extension of these dimension reduction approaches to multivariate nonlinear time series analysis has been shown to give good results in certain cases.

Chapter 3

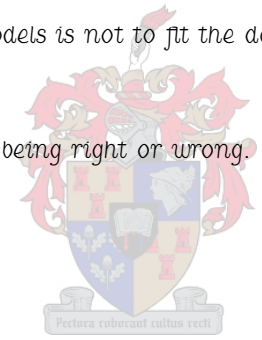
Nonlinear System Identification

"The purpose of models is not to fit the data but to sharpen the questions."

– Samuel Karlin

"A theory has only the alternative of being right or wrong. A model has a third possibility: it may be right, but irrelevant."

– Manfred Eigen



3.1 Introduction to Modelling

The objective in modelling to define a functional relationship which maps a given input space $\mathbf{X}$ to a corresponding output domain $\mathbf{Y}$:

$$y = f(\Phi(\mathbf{x})) \qquad f :\rightarrow \mathbb{R}^n \times \mathbb{R}^{n_o} \tag{3.1}$$

such that f will correctly approximate outputs y_i given future inputs $\mathbf{x}_i$. In particular, time series modelling seeks a predictive mathematical model $f : \mathbb{R}^{d_e} \rightarrow \mathbb{R}$ for the trajectory using mixtures of past measurements taken on the system. In other

words, given a set of observations, $\{x_t\}_{t=0}^{N}$, we seek to find a function of the form

$$\begin{aligned} x_{t+1} &= f(x_t, x_{t-T}, x_{t-2T}, \ldots, x_{t-(d_e-1)T}), \\ &= f(\mathbf{x}_t), \qquad \mathbf{x}_t \in \mathbb{R}^{d_e} \end{aligned} \tag{3.2}$$

for all $t = (d_e - 1)T + 1, \ldots, N - 1$, where d_e is the embedding dimension and T the time delay. Geometrically, equation (3.2) defines a d_e-dimensional surface or manifold in R^{d_e+1}. The embedding vectors $\mathbf{x}_t$ lie on a subspace in the embedding space R^{d_e} and the observation pairs $\{x_{t+1}, \mathbf{x}_t\}_{t=(d_e-1)T+1}^{N-1}$ lie on the manifold generated by $f \subset \mathbb{R}^{d_e+1}$. The *system identification* problem is to find a close enough approximation $\hat{f}$ to the true underlying function f,

$$\hat{x}_{t-1} = \hat{f}(\mathbf{x}_t) + e_t \tag{3.3}$$

where $\mathbf{x}$ is the state space and $e_t = x_{t-1} - \hat{x}_{t-1}$ is the discrepancy between actual and predicted values. Accordingly, the mathematical model described by equation (3.3) is sometimes called a state space model. Reconstruction of the system using only observed time series data facilitate the identification of the underlying dynamical system generating the observations (Packard *et al.*, 1980).

In *supervised learning* a state space mapping function f is found by using previous input–output pairs, called the *training data*, to train the learning machine or model. The input–output pairs are assumed to be generated according to some unknown probability distribution $P(\mathbf{x}, y)$. The target function f is the one that minimizes the expected error or *risk* given by

$$R[f] = \int \ell(f(\mathbf{x}), y) dP(\mathbf{x}, y) \tag{3.4}$$

where ℓ is an appropriately chosen *loss function* (see Cristianini and Shawe-Taylor, 2000; Müller *et al.*, 2001). One commonly used loss function in function approximation problems is the squared loss: $\ell(f(\mathbf{x}, y) = (f(\mathbf{x}, x) - y)^2$. Since the underlying

probability distribution $P(\mathbf{x}, y)$ is unknown, our best hope is an estimate of the target function f called the solution function $\hat{f}$ that is close to the otherwise *unknown* optimal using the available information, that is, the training set and properties of the function class or hypotheses space $\mathcal{F}$ from which the hypothesis or function f is chosen.

Using a suitable induction principle, such an appropriate $\hat{f}$ can be found, of which minimization of the empirical risk is one of the possible approaches:

$$R_{emp}[f] = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i, y_i) \tag{3.5}$$

Under certain conditions imposed on the learning machine, $R_{emp}[f]$ converges to $R[f]$ as $n \to \infty$.

Whilst it is preferable to minimize $\ell(f(\mathbf{x}, y)$ complications arise due to limited data of poor quality available. A fit with the least possible minimal error potentially explains peculiarities that are in fact due to uncertainties in the data rather than system dynamics, a phenomenon known as *overfitting*. Overfitting results in large deviations between future system outputs and predicted outputs.

It is therefore necessary to obtain a model fit that captures the dynamics correctly without including fluctuations introduced by measurement errors. In other words, a model must possess good generalization characteristics when presented with future inputs. This is referred to as the principle of parsimony – the simplest model or hypothesis in accordance with the observations should be preferred. This is also referred to as Occam's razor principle[1]. Generalization can be obtained by restricting the complexity of the function class $\mathcal{F}$ from which the estimate to the target function f is selected from. The need for restricting model complexity introduces the problem of model selection in the learning procedure.

[1] "pluralitas non est ponenda sine necessitate" – knowledge should only be obtained from observation, logical necessity, or divine revelation, and things not known to exist should not be postulated as existing, unless absolutely necessary.

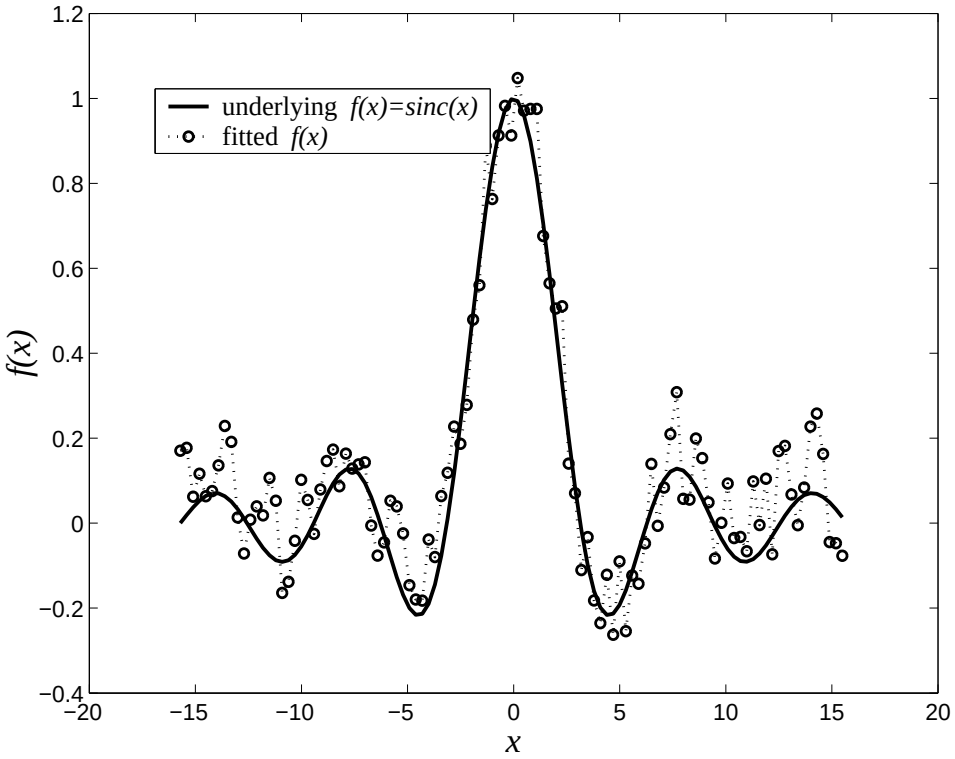


Figure 3.1: An example of overfitting during supervised learning. Instead of generalizing, the fitted function traces *exactly* the variations in the observed data.

Different model selection criteria exist that consider quantities such as the number of parameters, prediction error, etc. (Small, 1998). Figure 3.2 illustrates the general idea of how a compromise between model order (number of parameters) and model accuracy (prediction errors) can be obtained. The underlying principle in most of the criteria is to introduce a regularization term which penalizes the complexity of the selected function is selected. Three commonly used criteria are;

- Information Theory inspired criteria including Akaike Information Criterion (AIC), Schwarz (Bayesian) Information Criterion (SIC) (Judd and Mees, 1998).
- Rissanen's Minimum Description Length (MDL) Principle. The MDL principle proposes using the hypotheses space which compresses the data best,

that is, the description of the selected function and the list of corresponding training errors is shortest (Judd and Mees, 1998; Rissanen, 1999).

- Statistical Bounds on the generalization error based on the VC theory and structural risk minimization (SRM) principle (Cristianini and Shawe-Taylor, 2000). SRM selects a hypotheses space $\mathcal{F}_i$ (and function f_i) such that an upper bound on the generalization error is minimized;

$$R[f] \leq R_{emp}[f] + \varrho(h)$$

where h is the VC dimension and $\varrho(h)$ is the generalization term.

We are interested in nonlinear system identification based on the reconstruction of the state space of the system. A state space model ensures the underlying system dynamics are captured to allow for the prediction of the future time outputs. There are a number of possible function classes $\mathcal{F}$ that f can be chosen from. We distinguish between linear and nonlinear modelling by specifying a model is nonlinear if a nonlinear transformation is applied to any of the elements in the state space vector. For completeness, a brief description of a popular linear modelling technique is described in the following. Later, we compare results from using such an approach to nonlinear modelling techniques.

3.2 Linear Autoregressive Modelling

In autoregressive modelling (AR) it is assumed the process behaviour fluctuates around some fixed point. For zero-mean centered data, the fixed point is always zero. The linear autoregressive model gives an estimate of the current state x_t as a linear combination of the past m values

$$x_t = \sum_{k=1}^{m} a_k x_{t-k} + e_t \tag{3.6}$$

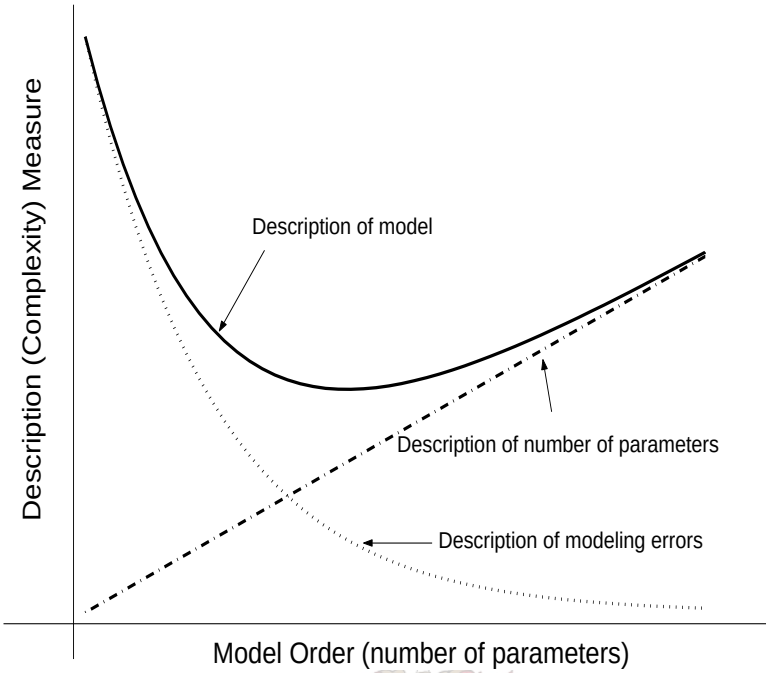


Figure 3.2: A schematic illustration of the variation of model accuracy with model complexity. As model accuracy increases (decreasing error) the number of parameters increases, which increases the model complexity. The more complex a model the poorer its generalization ability. An optimal model description seeks a balance between model complexity and model accuracy

where e_t are the residuals. The maximum likelihood approach finds estimates of the parameters a_k, $k = 1, ..., m$ by minimizing the sums of squares of e_t. The issue to be addressed then is the choice of model-order m.

3.3 Nonlinear Modelling

(a) ***Local Approximation Methods***

Local linear approaches form a predictor using behaviour of similar patterns in previously observed behaviour. Data points influence the function fit only

in a restricted or local neighbourhood and have no effect in the rest of the embedding space. A number of approaches are possible and are summarized as:

(i) Method of Analogues

From the reconstructed states, a segment of length d_e from the past that is most similar to the current point $[x_t, x_{t-T}, \ldots, x_{t-(d_e-1)T}]$ is identified and the data point following the identified segment is used as the predicted value, $\hat{x}_{t+1}$. This method is roughly equivalent to constructing a look-up table of previous states visited by the trajectory. The limitation of the approach, however, is it results in a future evolution exhibiting periodic oscillation.

(ii) Weighted Linear Combinations

This is an improvement on the method of analogues. One takes a linear combination of the nearest neighbours and uses the weighted average of their state mappings for predicting the future value.

The general mathematical formulation of local linear approximation methods described in (i) and (ii) is represented as

$$x_t = \sum_{k=1}^{m} f(\mathbf{x}_{t(k)}) \Phi(\| \mathbf{x}_t - \mathbf{x}_{t(k)} \|) \tag{3.7}$$

where $\mathbf{x}_{t(k)}$ denotes the k^{th} closest vector to $\mathbf{x}_t \in \mathbb{R}^m$ and k is the number of nearest neighbours. It is usual to take Φ as a *fixed* weight function increasing from zero to one when $\mathbf{x}_t$ approaches $\mathbf{x}_{t(k)}$. However, the approximating maps are generally not continuously differentiable, and the search for neighbours becomes time consuming as the number of vectors stored increase.

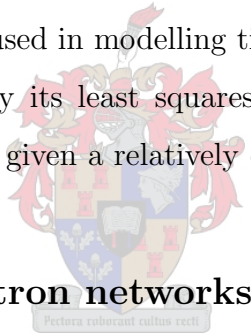
(iii) Simplicial

This class of linear approximation methods fits a surface in graph space

$\mathbb{R}^{m+1}$ to the observation pairs $(x_t, \mathbf{x}_t)$. Taking $k > m+1$ and fitting a plane we obtain a local autoregressive (AR) model or a local linear model (Lillekjendlie *et al.*, 1995).

(b) ***Global Approximation Methods***

In global model fitting all the data is used in defining a global function for the reconstructed attractor. A plethora of models classes exist from which $\hat{f}$ can be estimated. These include polynomial basis functions, rational approximations, Volterra function expansions, radial basis functions (RBF), multilayer perceptron neural networks (MLP), support vector machines (SVM), etc. The concepts and theory behind MLPs and SVMs are considered in some details next as they were primarily used in modelling time series data in this report. SVM learning, and especially its least squares extension, are still in early beginnings and hence will be given a relatively detailed account.



3.3.1 Multilayer perceptron networks (MLP)

Multilayer perceptrons are a subset of a powerful model class used in function approximation and pattern recognition called artificial neural networks. The neural network learning paradigm was inspired by the brain metaphor from cognitive sciences. The brain is a massive, parallel distributed structure constituted of structural units called *neurons*. The parallel distributed structure enables the brain to perform certain functions much faster than a digital computer. Furthermore, such a structure allows the brain to perform required tasks without expending much energy. This is in sharp contrast to a modern digital computer that uses lots of energy to perform even the simplest of tasks. An artificial neural network is designed to *mimic* how the brain performs its tasks. Neural networks used in a *learning* context employ a massive interconnection of structural units called nodes or neurons.

Haykin (1994) defines the neural network as follows:

A neural network is a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in that: (i) Knowledge is acquired by the network through a learning process. (ii) Interneuron connection strengths, or synaptic weights are used to store the knowledge.

Multilayer perceptron (MLP) networks have been used in nonlinear system identification (Lillekjendlie *et al.*, 1995). An MLP structure (Figure 3.3) is specified in terms of the model class, the network structure, and the basis function for each layer. Interconnected nodes are arranged in layers in the network. The MLP has three substructures; an input layer, one or more hidden layers, and an output layer. However, the input layer is generally not a true layer and, hence it is common to refer to the structure in terms of the number of hidden layers and the output layer. For example, an MLP with one hidden layer is called a two-layered network. Weights are assigned to the "synaptic" connections that forward transfer the outputs of nodes from each of the layers. It is also common to include a bias constant term per layer. The weights and biases then constitute the function parameters to be estimated by a supervised training algorithm. An input signal is propagated forward through all the layers in the network. MLPs iteratively adjust the parametric weights w_{ij} by supervised training using a *error back-propagation* algorithm, which is a generalization of the least squares algorithm. The error back-propagation algorithm proceeds in two passes. In the *forward pass* an input vector is applied to the input sensory nodes and its effect is propagated through the network to produce a set of outputs as the response of the network. The synaptic *weights* of the network are all fixed during the forward pass. In the *backward pass*, an error signal derived by subtracting the actual response of the network from the target signal, is back-propagated through the network, adjusting the synaptic weights according to some

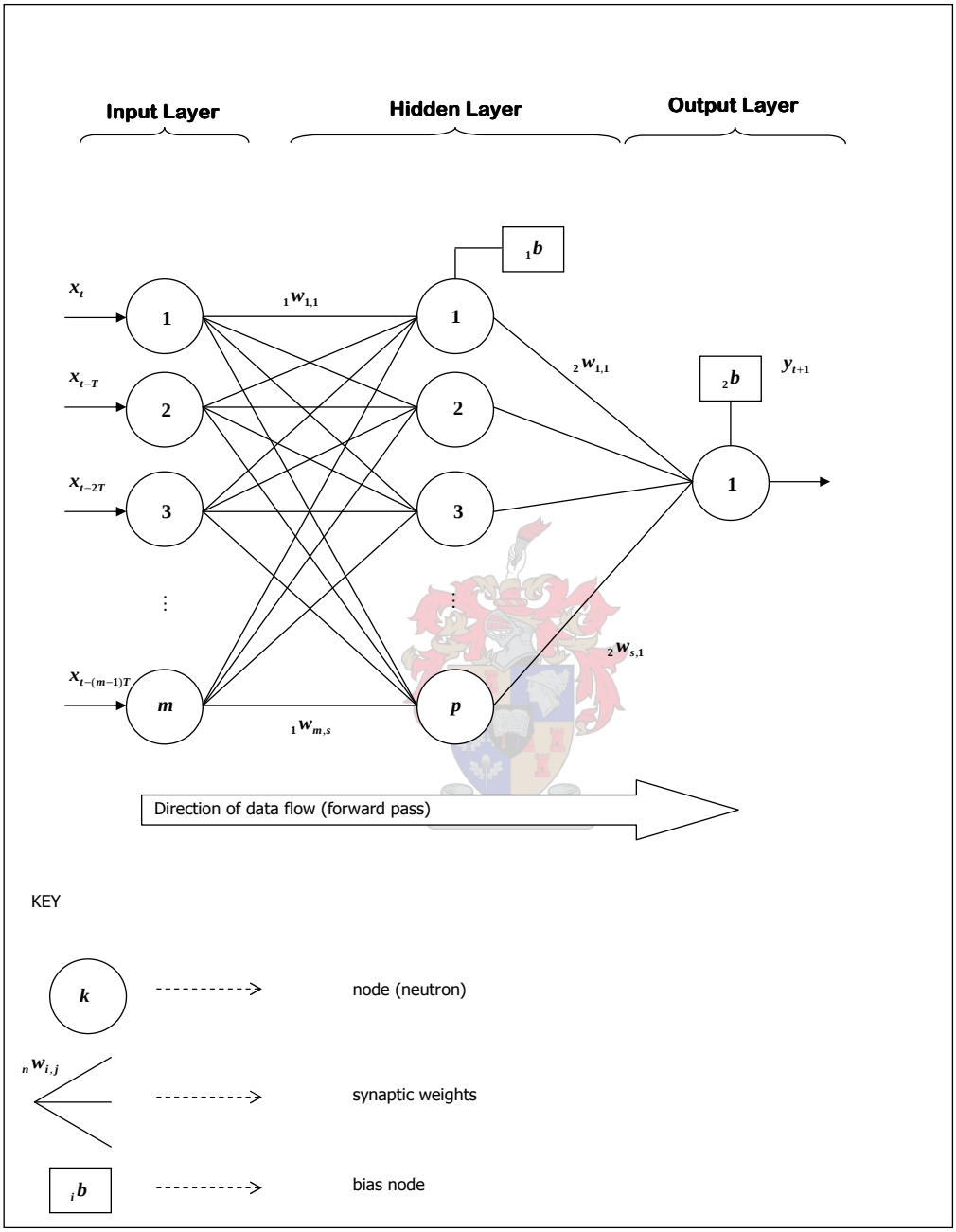


Figure 3.3: Multilayer perceptron network structure. ${}_kw_{i,j}$ are the synaptic weights, ${}_kb$ bias terms

error-criterion rule. A smooth nonlinearity is introduced by use of appropriate basis functions in the nodes of the hidden layer (and in the output layer, though this is rare). The most commonly used basis functions are the sigmoidal functions;

$$\phi(x) = \frac{1}{1+e^{-x}} \tag{3.8}$$

$$\phi(x) = \frac{e^{2x}-1}{e^{2x}+1} \tag{3.9}$$

The MLP can estimate almost any nonlinear function. However, MLPs are prone to be trapped in local minima of the error surface of the cost function used in the training of the network using gradient-descent algorithms, Figure 3.4.

Another drawback of MLPs is the extended parameter estimation time required, especially for complicated network structures. Proposals have been implemented that minimize or eliminate these disadvantages. For example, including a momentum term to speed up convergence and simulated annealing to avoid local minima. Despite these improvements, the MLP cannot provide a solution to all possible problems (Haykin, 1994).

If we consider a system with three input variables, x_t, x_{t-T}, x_{t-2T}, three hidden nodes with basis functions ϕ and a single output node, the MLP network defines a mapping $f : \mathbb{R}^3 \to \mathbb{R}$ given by

$$\begin{aligned} x_{t+1} &= \hat{f}(x_t, x_{t-T}, x_{t-2T}) \\ &= \phi_1\left(\sum_{k=1}^{3} {}_2w_{k,1}\phi_2\left({}_1w_{1,k}x_t + {}_1w_{2,k}x_{t-T} + {}_1w_{3,k}x_{t-2T}\right)\right) \end{aligned} \tag{3.10}$$

Details of the MLP concept and learning algorithm can be found in standard neural networks textbooks such as Haykin (1994).

3.3.2 Support vector machines (SVM)

The support vector machine (SVM) is a kernel-based learning algorithm based on statistical learning theory (also called VC theory) developed by mainly by Vapnik

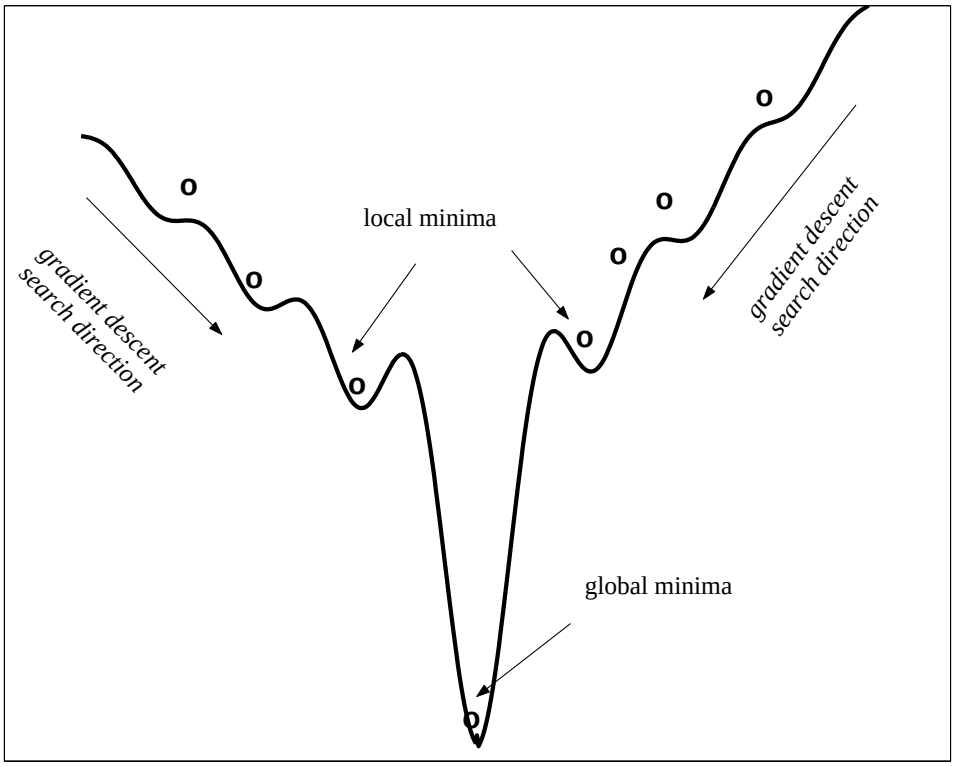


Figure 3.4: Local minima traps encountered in MLP training using gradient descent algorithms. The two local minima sites as indicated by the arrows risk in the MLP failing to reach the global minimum, resulting in poor performance of the trained model.

and Chervonenkis (Cristianini and Shawe-Taylor, 2000; Müller *et al.*, 2001; Smola, 1998). It is still a relatively new learning machine compared to other classes of learning machines. SVM solutions are obtained by solving a convex optimization problem. The model complexity follows from solving this convex optimization problem. The appeal to SVM models is that they scale very well to high-dimensional input spaces.

The SVM regression approach maps the input data $\mathbf{x}$ into a high-dimensional (possibly infinite) space Γ via a nonlinear mapping φ, facilitating for linear regres-

sion to be performed in this space;

$$f(\mathbf{x}) = \mathbf{w} \cdot \varphi(\mathbf{x}) + b, \qquad \varphi : \mathbb{R} \rightarrow \Gamma \tag{3.11}$$

where b is a threshold. Thus, linear regression in a high-dimensional (or feature) space corresponds to nonlinear regression in the low-dimensional input space $\mathbb{R}$, as illustrated in Figure 3.5(a).

Since Φ is fixed, $\mathbf{w}$ is determined from the data by minimizing the sum of the empirical risk and a complexity term;

$$R_{reg}[f] = R_{emp}[f] + \lambda \| \mathbf{w} \|_2^2 \tag{3.12}$$

$$= \sum_{i=1}^{N} \ell(\mathbf{x}, y) + \lambda \| \mathbf{w} \|_2^2 \tag{3.13}$$

where N is sample size, $\ell(\mathbf{x}, y)$ the cost function, λ a regularization constant, and $\| \cdot \|_2$ is the Euclidean norm. Equation (3.13) can minimized for a large set of functions by solving a quadratic programming problem for which a *unique* solution exists. One possible strategy is to keep the empirical risk zero by constraining $\mathbf{w}$ and b in equation (3.11) to the perfect fit case, while minimizing the complexity term, that is $\| \mathbf{w} \|_2^2$. (The Euclidean norm will be assumed for all norm definitions hereafter unless otherwise specified, hence the subscript will be dropped on the norm expression.)

Vapnik's ϵ-insensitive SV regression

Here, the objective is to fit a function f that has at most ϵ deviation from the targets for all points in the training set. In empirical risk minimization one optimizes the cost function

$$R_{emp} = \frac{1}{N} \sum_{i=1}^{N} |y_i - f(\mathbf{x}) - b|_\epsilon \tag{3.14}$$

where

$$|y - f(\mathbf{x})|_\epsilon = \begin{cases} 0 & \text{if } |y - f(\mathbf{x})| \leq \epsilon \\ |y - f(\mathbf{x})| - \epsilon, & \text{otherwise.} \end{cases} \tag{3.15}$$

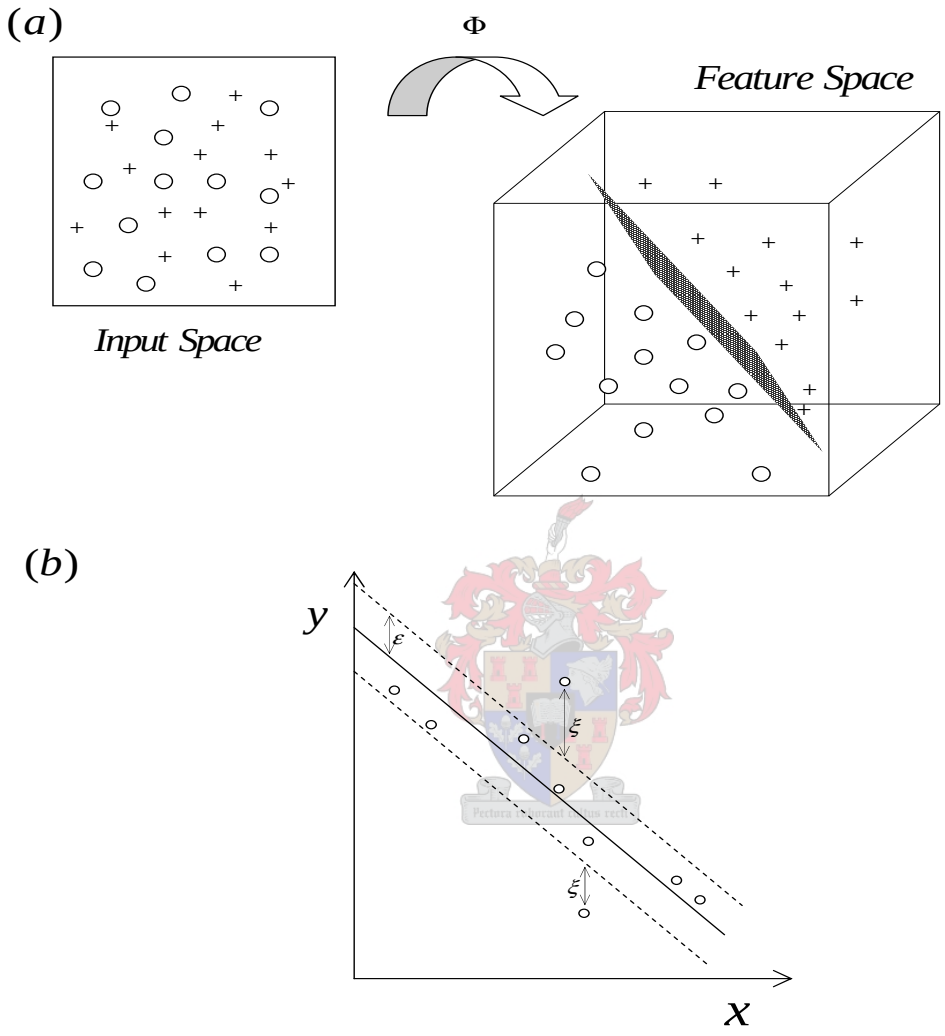


Figure 3.5: (a) The basic idea in SVM learning: Find a mapping Φ that transforms input data into a higher dimensional feature space in which linear separation is possible. (b) Vapnik's ϵ-insensitive band for one-dimensional linear regression problem

$|y - f(\mathbf{x})|_\epsilon$ is called Vapnik's ϵ-insensitive loss function (Cristianini and Shawe-Taylor, 2000). The corresponding convex optimization problem is;

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\| \mathbf{w} \|^2 \tag{3.16}$$

$$\text{subject to} \quad \begin{array}{l} y - (\mathbf{w} \cdot \mathbf{x}_i - b) \leq \epsilon \\ (\mathbf{w} \cdot \mathbf{x}_i + b) - y \leq \epsilon \end{array} \tag{3.17}$$

The required accuracy of the approximation is specified by ϵ. Slack variables ξ_i, ξ_i^* can be introduced to account for otherwise infeasible constraints in the preceding optimization scheme to yield;

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\| \mathbf{w} \|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \tag{3.18}$$

$$\text{subject to} \quad \begin{array}{l} y - (\mathbf{w} \cdot \mathbf{x}_i - b) \leq \epsilon + \xi_i \\ (\mathbf{w} \cdot \mathbf{x}_i + b) - y \leq \epsilon + \xi_i^* \end{array} \tag{3.19}$$

The constant C determines the trade-off between the smoothness (or complexity) of f and the tolerance level up to which deviations larger than ξ are tolerated.

Solving for $\mathbf{w}$ in Equations (3.17) and (3.19) directly means accessing the possibly infinite dimensional feature space, a task that demands excessive computational overhead. Fortunately, explicit usage of $\mathbf{w}$ can avoided by formulating the dual optimization problem. We re-formulate the optimization problem in terms of a Lagrangian function defined from the objective function and the corresponding constraints and introducing a dual set of variables $\alpha_i, \alpha_i^* \geq 0, i = 1, \ldots, n$. The Lagrange formulation for the optimization scheme in equations (3.19) is

$$\begin{array}{rcl} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*) & = & \frac{1}{2}\| \mathbf{w} \|^2 - C \sum_i^n (\xi_i + \xi_i^*) \\ & & - \sum_{i=1}^N \alpha \{(y - (\mathbf{w} \cdot \mathbf{x}_i - b - \epsilon - \xi_i)\} \\ & & - \sum_{i=1}^N \alpha^* (\mathbf{w} \cdot \mathbf{x}_i + b - y - \epsilon - \xi_i^*) \\ & & - \sum_{i=1}^N \nu_i \xi_i - \sum_{i=1}^N \nu_i^* \xi_i^* \end{array} \tag{3.20}$$

The corresponding dual formulation is found by differentiating L with respect to $\mathbf{x}, \xi, \xi^*$, and b and setting the resulting equations to zero, that is;

$$\begin{aligned}
\frac{\partial L(\mathbf{w}, b, \xi, \alpha, \alpha^*)}{\partial \mathbf{w}} &= 0 \\
\frac{\partial L(\mathbf{w}, b, \xi, \alpha, \alpha^*)}{\partial b} &= 0 \\
\frac{\partial L(\mathbf{w}, b, \xi, \alpha, \alpha^*)}{\partial \xi} &= 0 \\
\frac{\partial L(\mathbf{w}, b, \xi, \alpha, \alpha^*)}{\partial \xi^*} &= 0
\end{aligned} \tag{3.21}$$

Solving the equations (3.21) and substituting for $\mathbf{w}$ and b in equation (3.20) reduces the dual optimization problem to;

$$\begin{aligned}
\max_{\alpha,\alpha^*} \quad & -\tfrac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_j^*)(\alpha_j - \alpha_j^*)(\mathbf{x}_i \cdot \mathbf{x}_j) \\
& +\varepsilon \sum_{i,j=1}^{N}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{N} y_i(\alpha_j - \alpha_j^*) \\
\text{subject to} \quad & \sum_{i}^{N}(\alpha_i - \alpha_i^*) = 0 \\
& 0 \leq \alpha, \alpha_i^*, \leq C
\end{aligned} \tag{3.22}$$

The decision rule then becomes

$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)(\mathbf{x}_i \cdot \mathbf{x}) + b \tag{3.23}$$

Equation (3.23) is the standard SVM algorithm and the expression is called the *support vector expansion* as $\mathbf{w}$ is completely described by a linear combination of the training patterns $\mathbf{x}_i$. The standard SVM algorithm only depends on the dot products between various patterns in the input space and is defined for linear problems. For nonlinear problems, the input space is mapped into a high-dimensional feature space $\Phi : \mathbb{R}^n \rightarrow \Gamma$ and the decision rule is then expressed as

$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)(\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) + b \tag{3.24}$$

Use of kernel representation, $K(\mathbf{x}, \mathbf{y}) \equiv \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})$, allows the application of the standard SV algorithm without explicit knowledge of what the feature space is as

long as the kernel used fulfills Mercer's condition. Figure 3.6 illustrates the use of the kernel trick in transforming a nonlinear problem into a high-dimensional space where linear separation is possible. The kernel trick is useful in that one does not need to know what this high-dimensional is (Cristianini and Shawe-Taylor, 2000; Smola, 1998).

Table 3.1 lists some common kernel functions, $K(\mathbf{x}, \mathbf{y})$. For radial basis functions or sigmoidal kernels the number of hidden nodes corresponds to the number of support vectors.

Table 3.1: Common kernel functions

Name	Kernel Function, $K(\mathbf{x}, \mathbf{y}$
linear SVM	$\mathbf{x} \cdot \mathbf{y}$
Gaussian RBF	$\exp\left(\frac{-\parallel \mathbf{x} - \mathbf{y} \parallel^2}{\sigma^2}\right)$
Polynomial of degree d	$((\mathbf{x} \cdot \mathbf{y}) + \theta)^d$
Sigmoidal	$\tanh(\kappa(\mathbf{x} \cdot \mathbf{y}) + \theta)$
inv. multiquadric	$\frac{1}{\sqrt{\parallel \mathbf{x} - \mathbf{y} \parallel^2 + c^2}}$

Using kernel representation, the nonlinear support vector algorithm corresponding to equation (3.22) is

$$\begin{aligned} \max_{\alpha,\alpha^*} \quad & -\tfrac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_j^*)(\alpha_j - \alpha_j^*)K(\mathbf{x}_i \cdot \mathbf{x}_j) \\ & +\varepsilon\sum_{i,j=1}^{N}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{N} y_i(\alpha_j - \alpha_j^*) \\ \text{subject to} \quad & \sum_{i}^{N}(\alpha_i - \alpha_i^*) = 0 \\ & 0 \leq \alpha, \alpha_i^*, \leq C \end{aligned} \tag{3.25}$$

and the corresponding support vector expansion is

$$\parallel \mathbf{w} \parallel \;\; = \;\; \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\varphi(\mathbf{x}_i) \tag{3.26}$$

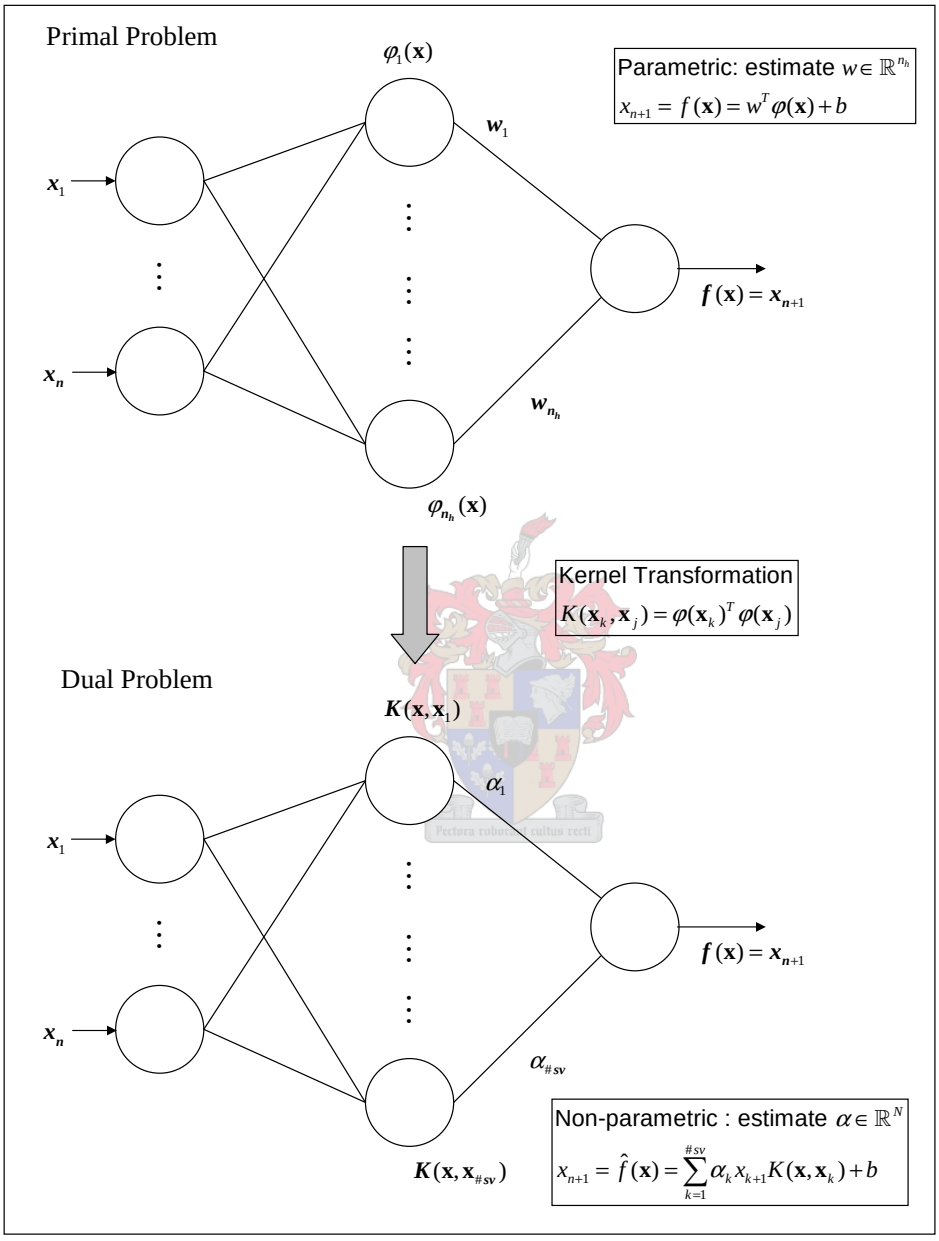


Figure 3.6: The primal-dual representation of SVMs. Duality is achieved through the use of a kernel transformation

$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)K(\mathbf{x}_i \cdot \mathbf{x}) + b \tag{3.27}$$

Least squares support vector machines (LSSVM)

Suykens *et al.* (2000) and Suykens (2001) proposed a modification of Vapnik's standard SVM by replacing the ϵ-insensitive loss function with a least-squares cost function. Furthermore, equality constraints instead of inequality constraints are considered in the problem formulation. Consequently, one solves a linear system instead of a quadratic programming problem. Thus, given the training set $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, the optimization problem is formulated as follows

$$\begin{aligned} \min_{\mathbf{w},b,e} \quad & \Im(\mathbf{w}, e) = \frac{1}{2}(\mathbf{w}^T\mathbf{w} + \lambda \sum_{k=1}^{N-1} e_k^2 \\ \text{subject to} \quad & y_k = \mathbf{w}^T\Phi(\mathbf{x}_k) + b + e_k \end{aligned} \tag{3.28}$$

The cost function with a squared error term and a regularization term λ corresponds to a form of ridge regression. The squared error term and equality constraints allow for simplification of the problem. The corresponding Lagrangian function can is

$$L(\mathbf{w}, b, e; \alpha) = \Im(\mathbf{w}, e) - \sum_{k=1}^{N} \alpha_k[(w)^T\Phi(\mathbf{x}_k + b + e_k - y) \tag{3.29}$$

where the α_k's are the Langrage multipliers. For optimality the following stationarity conditions must hold;

$$\begin{aligned} \frac{\partial L}{\mathbf{w}} = 0 \quad &\rightarrow \quad \mathbf{w} = \textstyle\sum_{k=1}^{N} \alpha_k \Phi(\mathbf{x}_k) \\ \frac{\partial L}{\partial b} = 0 \quad &\rightarrow \quad \textstyle\sum_{k=1}^{N} \alpha_k = 0 \\ \frac{\partial L}{\partial e_k} = 0 \quad &\rightarrow \quad \alpha_k = \lambda e_k, \qquad\qquad k = 1, \ldots, N-1 \\ \frac{\partial L}{\partial \alpha_k} = 0 \quad &\rightarrow \quad \mathbf{w}^T\Phi(\mathbf{x}_k) + b + e_k - y_k = \lambda e_k, \quad k = 1, \ldots, N-1 \end{aligned} \tag{3.30}$$

Defining

$$\begin{aligned}
\mathbf{Z} &= [\varphi(\mathbf{x}_0)^T y_0; \ldots ; \varphi(\mathbf{x}_N)^T y_N] \\
\mathbf{Y} &= [y_1; \ldots ; y_N] \\
\vec{1} &= [1; \ldots ; 1] \\
\mathbf{e} &= [e_1; \ldots ; e_N] \\
\alpha &= [\alpha_1; \ldots ; \alpha_N]
\end{aligned} \tag{3.31}$$

and eliminating $\mathbf{w}$ and $\mathbf{e}$ the solution of equation 3.29 is

$$\left[\begin{array}{cc} 0 & \vec{1}^T \\ \hline \vec{1}^T & \Omega + \gamma^{-1}\mathrm{I} \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \vec{1} \end{array}\right] \tag{3.32}$$

where $\Omega = ZZ^T$. Application of Mercer's condition to the Ω matrix gives

$$\begin{aligned}
\Omega_{kl} &= \Phi(\mathbf{x}_k) \cdot \Phi(\mathbf{x}_l), \qquad\qquad k, l = 1, \ldots , N \\
&= K(\mathbf{x}_k, \mathbf{x}_l)
\end{aligned} \tag{3.33}$$

The resulting LSSVM for function estimation becomes

$$y_k = \sum_{k=1}^{N} \alpha_k K(\mathbf{x}_l, \mathbf{x}_k) + b \tag{3.34}$$

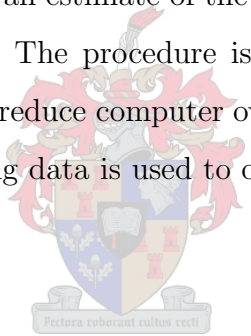
where α_k and b are the solution to the linear system.

However, use of the least squares norm results in loss of sparseness as all the support values are proportional to the training errors, equation (3.30). Suykens *et al.* (2000) have shown that by plotting the spectrum of the sorted $|\alpha_i|$ values one can evaluate those data points contributing significantly to the LSSVM model. Sparseness is then imposed by gradually omitting the least important data from the training set and re-estimating the model.

An important step in training the LSSVM is selection of the regularization parameter γ and kernel-related parameters, such as the kernel width σ for Gaussian RBF kernels. These parameters are referred to as hyperparameters. Hyperparameter selection can be done by minimizing either an estimate of generalization error

or some other related performance measure.Several performance measures exist and these include cross-validation, leave-one out methods, VC bounds approaches, and Bayesian learning methods (Duan *et al.*, 2001; Suykens *et al.*, 2000).

The k-fold cross-validation technique is one of the most robust methods (Duan *et al.*, 2001) and is used as the performance measure in this work. In k-fold cross-validation, the input space is assumed to independent and identically distributed. The training data is randomly split into k disjunct subsets (the folds) of approximately equal size. At each i-th iteration ($i = 1, \ldots, k$), the decision rule in equation (3.34) is obtained using $(k-1)$ subset as training data and the i-th fold is used to validate the trained model. After k iterations the average test error (usually mean square error) over the k-folds gives an estimate of the expected generalization error for the selected hyperparameters. The procedure is computationally demanding and modifications are necessary to reduce computer overhead time. In particular, a representative subset of the training data is used to optimize the hyperparameters as illustrated later in Chapter 4.



3.4 Evaluating Model Performance

There are different measures for evaluating the performance of the fitted model. We shall primarily use the mean square error. Given a model f that predicts future values according to $\hat{y}_i = f(\mathbf{x}_i)$, where $\mathbf{x}_i$ is a suitably chosen embedded vector, the mean square error (MSE) is the average of the square of the deviations between actual and predicted values and normalized by the standard deviation of the time series channel to be predicted σ_y;

$$MSE = \frac{\frac{1}{n}\sum_i^n (y_i - \hat{y}_i)^2}{\sigma_y} \tag{3.35}$$

Evaluation is done "honestly". In other words, a test set not used in building the model is used in assessing model performance. A MSE value of 0 indicates perfect

prediction and a value of 1 means that the model is as good as just using the average of the data for the forecasts. Hence, a close a value to 0 is sought for good predictive models.

The fitness of an estimated model is tested using two linear statistics, the R^2 statistic and the estimated correlation coefficient $\hat{\rho}$ defined by

$$R^2 = 1 - \frac{1}{(n-1)\sigma_y^2}\sum(y_i - \hat{y}_i)^2 \tag{3.36}$$

$$\hat{\rho} = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}} \tag{3.37}$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the mean of the actual and predicted time series, and the other variables as defined previously.

3.5 Concluding Remarks

Fitting a global model to input-output pairs involves the parameterization of suitably chosen model structure with known basis functions in a scheme known as supervised learning. Multilayer neural networks and support vector machines possess certain attributes that make them good candidates for nonlinear system identification. Use of either MLPs or SVMs involves minimizing a loss function on the training data whilst simultaneously choosing a structure or parameters that allow the network to have good generalization capabilities. The least-squares support vector machine is a variation of the SVM where one optimizes a linear instead of a quadratic function.

Chapter 4

Case Study: A Coupled CSTR System

"...Thus in the field of abstract thought the enquiring mind can never rest until it reaches the extremes, [where there is] a clash of forces freely operating and obedient to no law but their own" –Carl von Clausewitz

In this chapter the application of multivariate time series analysis will be explored using the coupled continuously stirred tank reactor (CSTR) system as a prototype signal generator. To assess the effectiveness of the multivariate approach different measures are used, including nonlinear predictive modeling, information theoreticals, and dynamical invariants. The performance, robustness, and generalization capabilities of a relatively new learning machine, the least-squares support vector machine, is compared with multilayer perceptron networks. The relative merits of principal component analysis and independent component analysis in dimensionality reduction are considered.

4.1 System Description

Cubic autocatalysis with catalyst decay of two isothermal reactions occuring in parallel in a CSTR have been studied widely, for example inLee and Chang (1996) and Abasher and Judd (1998). Lynch (1992a) generalized the CSTR system to include the possibility of higher reaction orders by adding a second autocatalytic step in the original Gray-Scott model (Gray and Scott, 1983) that made it possible to observe chaos in the reactor. Modeling of the coupled cubic auto-catalator studied previously by Abasher and Judd (1998) but from a synchronization perspective will be investigated in the following.

The reactions occurring in a cubic auto-catalator proceed according to

$$A + 2B \rightarrow 3B \qquad -r_A = k_1 C_A C_B^2 \tag{4.1}$$

$$D + 2B \rightarrow 3B \qquad -r_D = k_2 C_D C_B^2 \tag{4.2}$$

$$B \rightarrow C \qquad r_C = k_3 C_B \tag{4.3}$$

Coupling of two similar cubic auto-catalators by allowing bi-directional mass transfer between the systems (that is q_{ij} in Figure 4.1) affects the time evolution of each of the reactor's dynamics because of the effect of species transferred from either reactor. Material balance across the coupled system yields the following dimensionless differential equations

$$\frac{d}{d\tau}\begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \\ X_2 \\ Y_2 \\ Z_2 \end{bmatrix} = \begin{bmatrix} (1 - X_1) - Da_X X_1 Z_1^2 + \mu(X_2 - X_1) \\ \beta - Y_1 - Da_Y Y_1 Z_1^2 + \mu(Y_2 - Y_1) \\ 1 - (1 + Da_Z)Z_1 + \alpha(Da_X X_1 Z_1^2) + Da_Y Y_1 Z_1^2 + \mu(Z_2 - Z_1) \\ (1 - X_2) - Da_X X_2 Z_2^2 + \mu(X_1 - X_2) \\ \beta - Y_2 - Da_Y Y_2 Z_2^2 + \mu(Y_1 - Y_2) \\ 1 - (1 + Da_Z)Z_2 + \alpha(Da_X X_2 Z_2^2) + Da_Y Y_2 Z_2^2 + \mu(Z_1 - Z_2) \end{bmatrix} \tag{4.4}$$

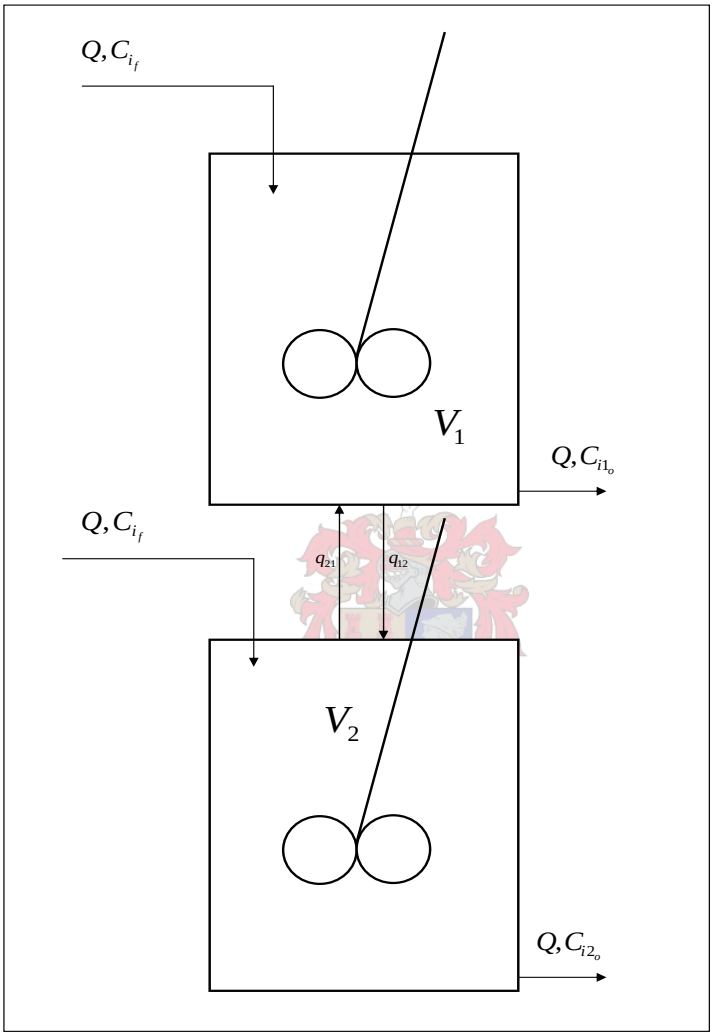


Figure 4.1: The coupled CSTR systems with a bi-directional mass transfer q between them.

where the ratios of species in the feed and dimensionless time by

$$\alpha = \frac{C_{A_f}}{C_{B_f}}, \quad \beta = \frac{C_{C_f}}{C_{A_f}}, \quad \mu = \frac{q}{Q}, \quad \tau = \frac{Qt}{V}$$

X_j, Y_j, Z_j are the dimensionless concentrations for a the respective species in reactor $j = \{1, 2\}$ given by

$$X_j = \frac{C_{A_j}}{C_{A_0}}, \quad Y_j = \frac{C_{D_j}}{C_{A_0}} = \beta \frac{D_j}{D_0}, \quad Z_j = \frac{C_{B_j}}{C_{B_0}}, \qquad j = \{1, 2\};$$

and the Damköhler numbers Da_X, Da_Y, Da_Z for species A, D and B respectively are defined by

$$Da_X = \frac{k_1 V C_{B_f}^2}{Q}, \quad Da_Y = \frac{k_2 V C_{B_f}^2}{Q}, \quad Da_Z = \frac{k_3 V}{Q}$$

For uncoupled case ($\mu = 0$), each of the CSTR exhibits chaotic behaviour for the parameter values reported by Lynch (1992a,b), namely, $\alpha = 1.5, \beta = 2.85, Da_X = 18000, Da_Y = 400, Da_Z = 80$. Abasher and Judd (1998) reported that coupling introduced more complicated behaviour than a single reactor for certain ranges of the coupling strengths as shown in Table 4.1.

Table 4.1: The dynamic behaviour exhibited by a coupled CSTR system in different coupling strength parameter zones.

Coupling	Description of dynamic behaviour
$0 < \mu < 0.885$	Hyperchaos, interrupted by periodic windows in the region $0.02 < \mu \leq 0.131$.
$\mu \geq 0.885$	Chaos (synchronization)

Synchronization of the two coupled CSTRs is achieved when the coupling strength is at least 0.885. The dynamic behaviour of the coupled CSTRs is asymptotically similar to a single uncoupled reactor. In other words, the reactors are practically uncoupled when synchronization is achieved[1].

[1]Boccaletti *et al.* (2002) define synchronization of chaos as a process wherein two (or many)

4.2 Data Generation

The system in equation (4.4) was simulated using the MATLAB® ode45 ordinary differential solver, which implements the Runge-Kutta fourth and fifth order formula, for a coupled CSTR system with parameter values given in Table 4.2.

Table 4.2: Parameter values used in numerical simulation of the coupled CSTR system

Parameter	Value
α	1.50
β	2.85
Da_X	18,000
Da_Y	400
Da_Z	80
μ	0.45
Integration time step $d\tau$	0.01
$(X_{1_0}, Y_{1_0}, Z_{1_0}, X_{2_0}, Y_{2_0}, Z_{2_0})$ (initial conditions)	(0.1,0.2,0.3,0.001,0.002,0.003)
Sampling interval	0.03τ
Integration range	[1 1500]
Transient steps	[1 500]

For certain values of the coupling strength (including $\mu = 0.45$) a random-like perturbation is induced on either of the individual systems. Consequently, in the absence of information on the coupling between the reactors, the dynamic behaviour of the species in each individual reactor will have a superimposed "noise" component as shown in Figure 4.3.

chaotic systems (either equivalent or nonequivalent) adjust a given property of their motion to a common behaviour, due to coupling or forcing.

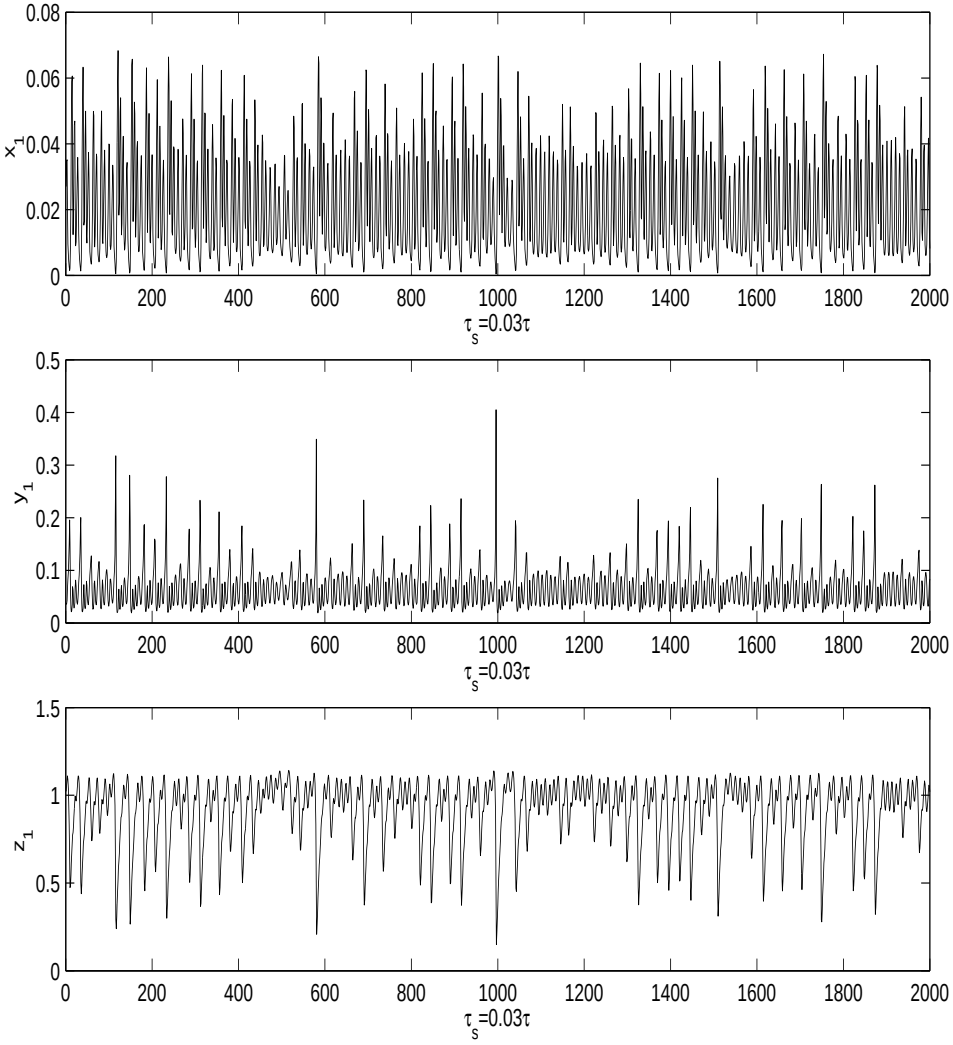


Figure 4.2: Time series plots of the dimensionless variables x_1, y_1, z_1 for coupling strength $\mu = 0.45$. Notice the irregular noise-like structure typically associated with stochastic signals. Similar plots are obtained for x_2, y_2, z_2 respectively.

Figure 4.4 are different projections of a typical attractor reconstructed from the original variables taken from a single reactor whilst Figure 4.5 is the underlying attractor when all the original variables are considered. It can be seen that the system is restricted to certain values for each of the species $x_j, y_j,$ and $z_j, j = \{1, 2\}$.

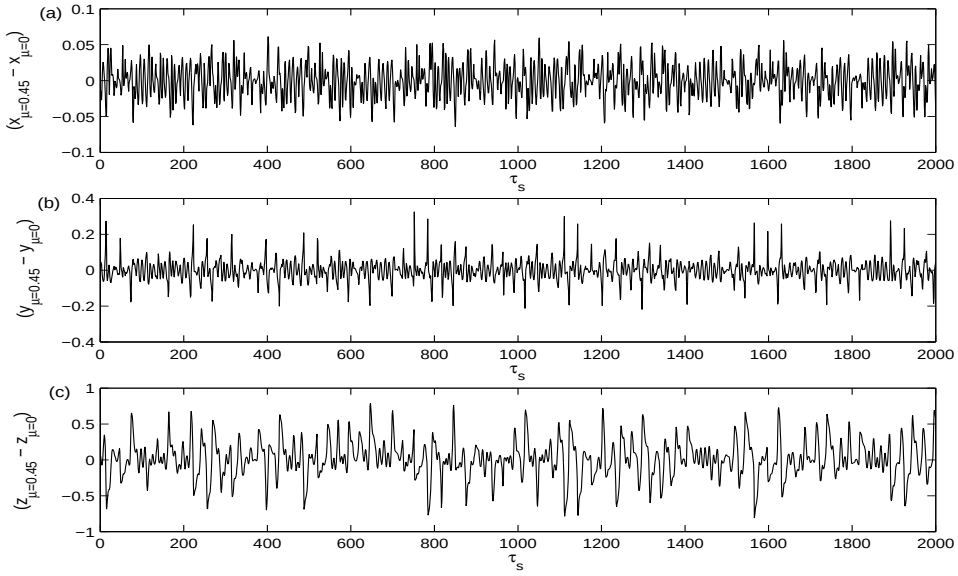


Figure 4.3: Effect of coupling on the reaction dynamics of reactor 1 dynamics. A similar effect is observed on reactor 2.

However, despite this restriction, the system continuously evolves in this space without repeating any segment of the trajectory. This behaviour is called *chaos* and the attractor is said to exhibit a *self-similar* or *fractal* structure.

Having generated the time series for each of the species in reactor 1 (and 2), it is henceforth assumed that the dynamical system generating the signals is unknown. The task, therefore, is to *reconstruct the system from a single or a combination of the simultaneously observed signals from the system.* Without loss of generality, in the following species x_1, y_1, z_1 are alternatively referred to as x, y, z respectively.

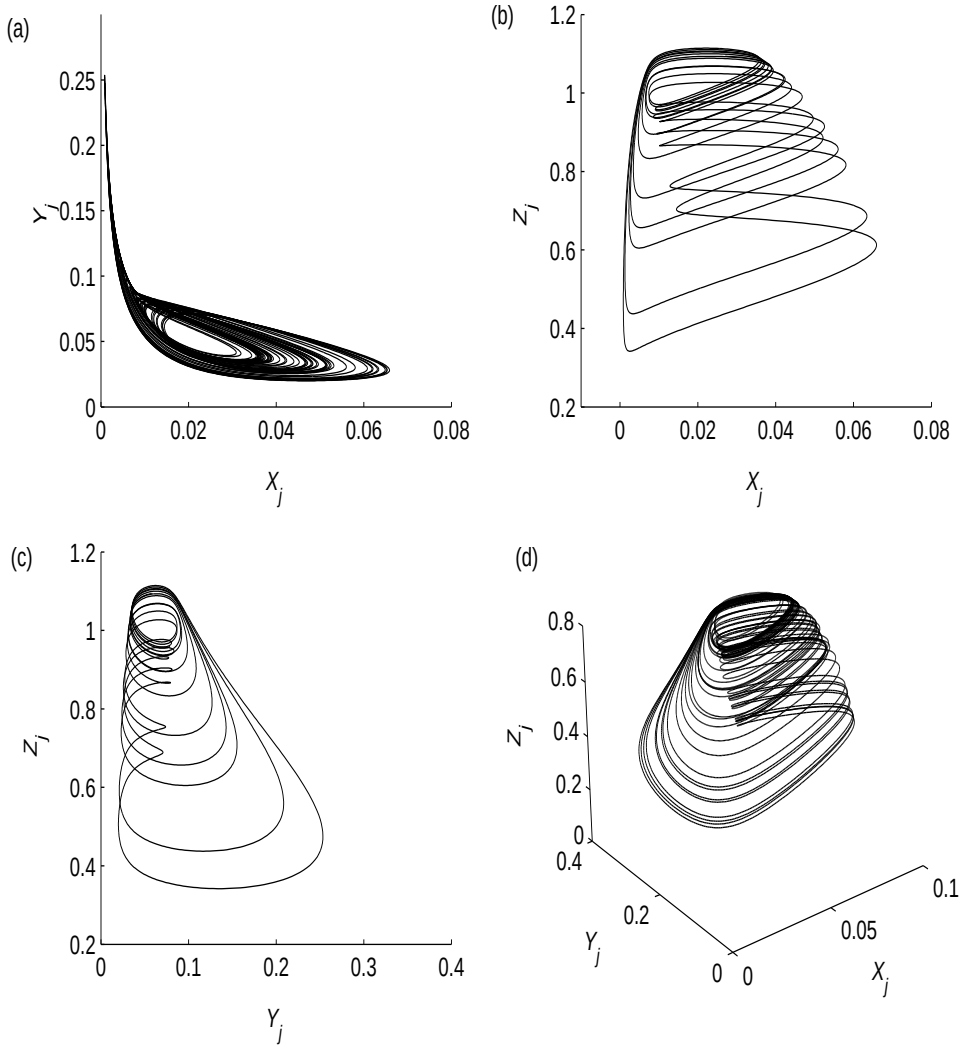


Figure 4.4: Projections of the reconstructed attractor of the coupled CSTR from a trivial embedding using original state variables from a single reactor $\in \mathbb{R}^3$.

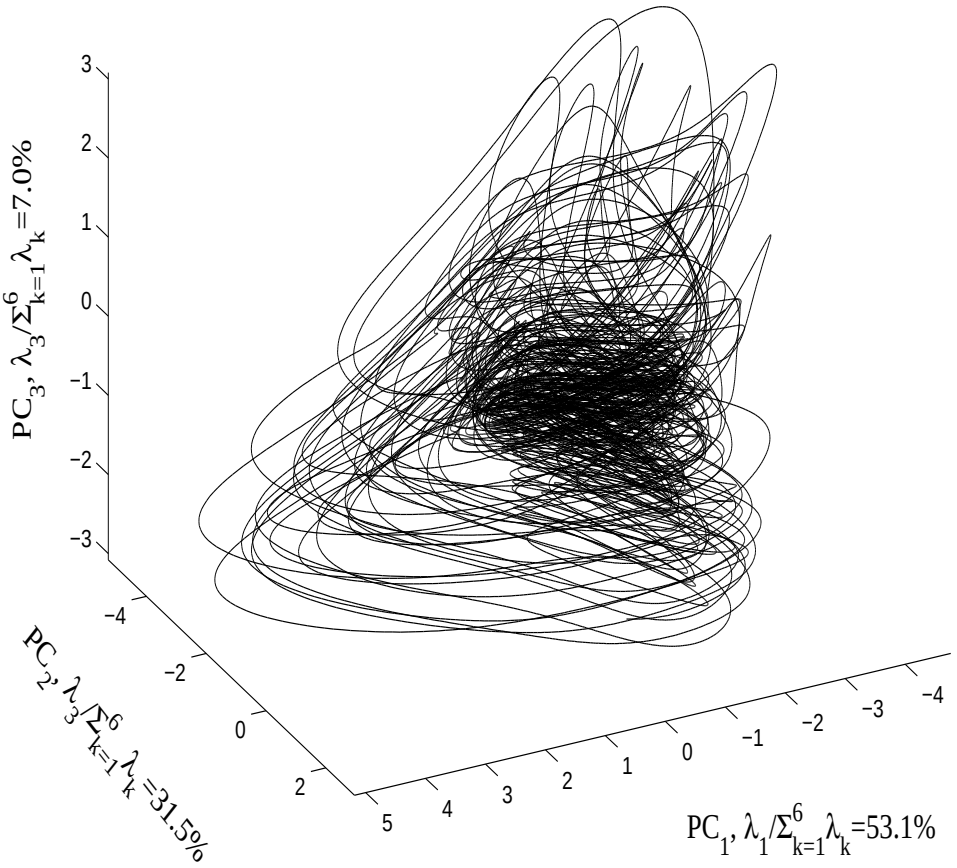


Figure 4.5: Original attractor reconstructed using the state variables $\in \mathbb{R}^6$ projected onto the three largest principal components that explain 91.6% of the total variance.

4.3 Determining the Embedding Parameters

As argued in Chapter 2, multivariate time series analysis is potentially superior than scalar time series. However, a good attractor reconstruction scheme is necessary to unfold the dynamics in phase space. Different approaches for determining the optimal embedding have been used by other researchers, for example Cao *et al.*

(1998) and Barnard *et al.* (2001). The approach of Barnard (1999) and Barnard *et al.* (2001) was used to define a multicomponent embedding. That is, individual components were separately embedded and then combined to give a multicomponent embedding, which was then optimally separated subspace using latent vector projection methods. In addition, the methods of Judd and Mees (1996, 1998), Judd *et al.* (1999), and Small (1998) that consider the embedding problem as a modeling problem were extended to the multivariate case.

Embedding or state space reconstruction makes it possible to study the dynamics of the possibly unknown original phase space. Given an infinite amount of data measured to an infinite resolution delay embedding theorems allow for the use of any time delay T. However, this is not possible in practice. Therefore, approximate values of the optimal delay that ensure a one to one mapping of the reconstructed attractor in phase space must be determined. Also, *any* value of the embedding dimension d_e ensures a deterministic mapping if it is sufficiently large (greater than the fractal dimension). However, in reconstructing dynamics from a noisy time series, arbitrarily large values of d_e result in every coordinate of the reconstructed phase space being populated by points generated by a random source. Also, high-dimensional space has an undesirable penalty as the computational cost and number of free parameters used in function approximation scale exponentially with dimension – *"curse of dimensionality"*. Therefore, an appropriate d_e optimized for the *specific* problem under investigation has to be determined.

Figures 4.7–4.9 are results of the embedding parameters obtained for each of the observed components $x, y,$ and z, determined using various methods as stated. Table 4.3 summarizes the estimates of the delay lag T and the embedding dimension d_e. The values of T suggested by the autocorrelation function and mutual information are not very different. Since the method of embedding is supposed to be insensitive to small differences in the T used (Kantz and Schreiber, 1997), either of the values

can be used. However, the values of d_e found using the false nearest neighbours algorithm and Cao's method differed significantly. In all cases the FNN method suggested a lower value. As explained in the Chapter 2, the FNN implementation is sensitive to the length of data used. Also, the case of d_e for y was ambiguous as the plot initially decreased to 3 and then fluctuated before decreasing to 10. The decay of the curve in the FNN plots was not due saturation of d_e values but the lack of enough data points required to resolve the unfolding of the attractor. Such ambiguity does arise in the implementation of the false nearest neighbours (Cao, 1997). For these reasons the d_e values determined using Cao's method were used in the reconstruction of phase space . Figure 4.6 is the reconstructed attractor from

Table 4.3: Summary of estimates of T using the autocorrelation function and Shannon's average mutual information, and the embedding dimension d_e using the false nearest neighbours algorithm and Cao's minimum embedding dimension criteria.

	Delay Lag T		Embedding Dimension d_e	
Variable	C_T	I_T	FNN	Cao
x	2	3	3	11
y	2	2	(3)10	13
z	5	4	6	11

Table 4.4: Determining non–uniform embedding lag vector using autoregressive modeling and the Rissanen's MDL principle. Choice of a suitable lag vector effectively specifies the embedding dimension, d_e.

Variable	Embedding lag vector, l	Effective d_e
x	$[0, 1, 2, 3, 5, 6, 8, 12, 15, 21, 24, 26] \leftarrow l_x$	12
y	$[0, 1, 4, 8, 10, 13, 15, 19] \leftarrow l_y$	8
z	$[0, 1, 3, 10] \leftarrow l_z$	4

the embedding of the x variable data points with $(d_e, T) = (11, 3)$ as determined above. The structure is visually different from the original attractor in Figure 4.5. However, Takens' theorem guarantees that analysis of the reconstructed attractor yields similar results as the original attractor.

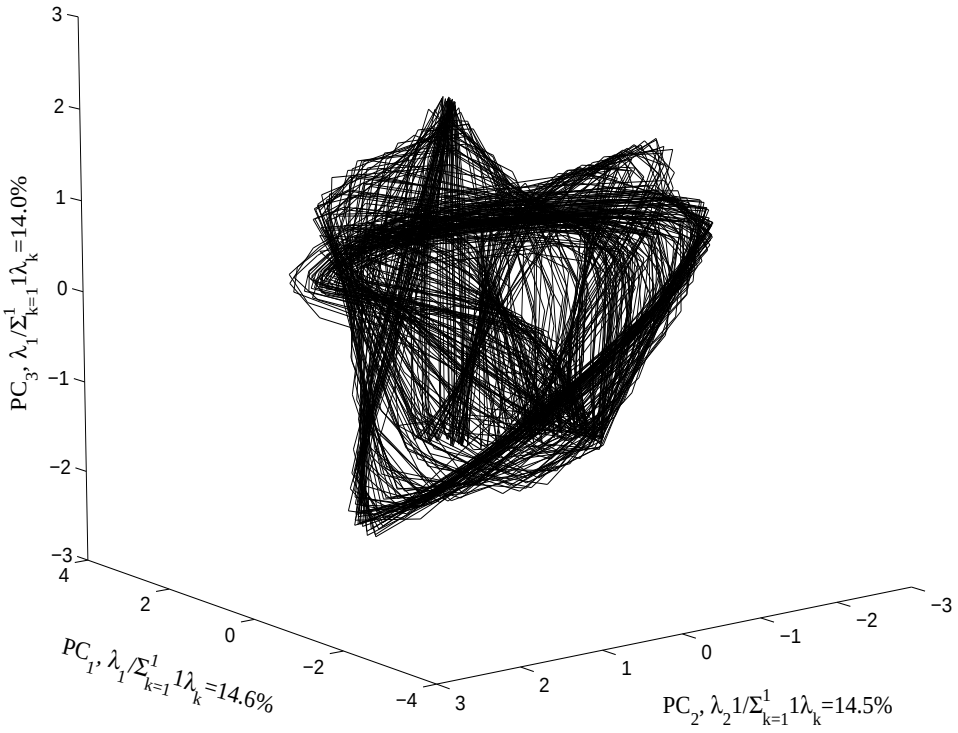


Figure 4.6: The reconstructed attractor $\subset \mathbb{R}^{11}$ of the coupled CSTR using x_1, with $(d_e, T) = (11, 3)$ projected onto the first 3 principal components that capture 43% of the variance in the multidimensional embedding.

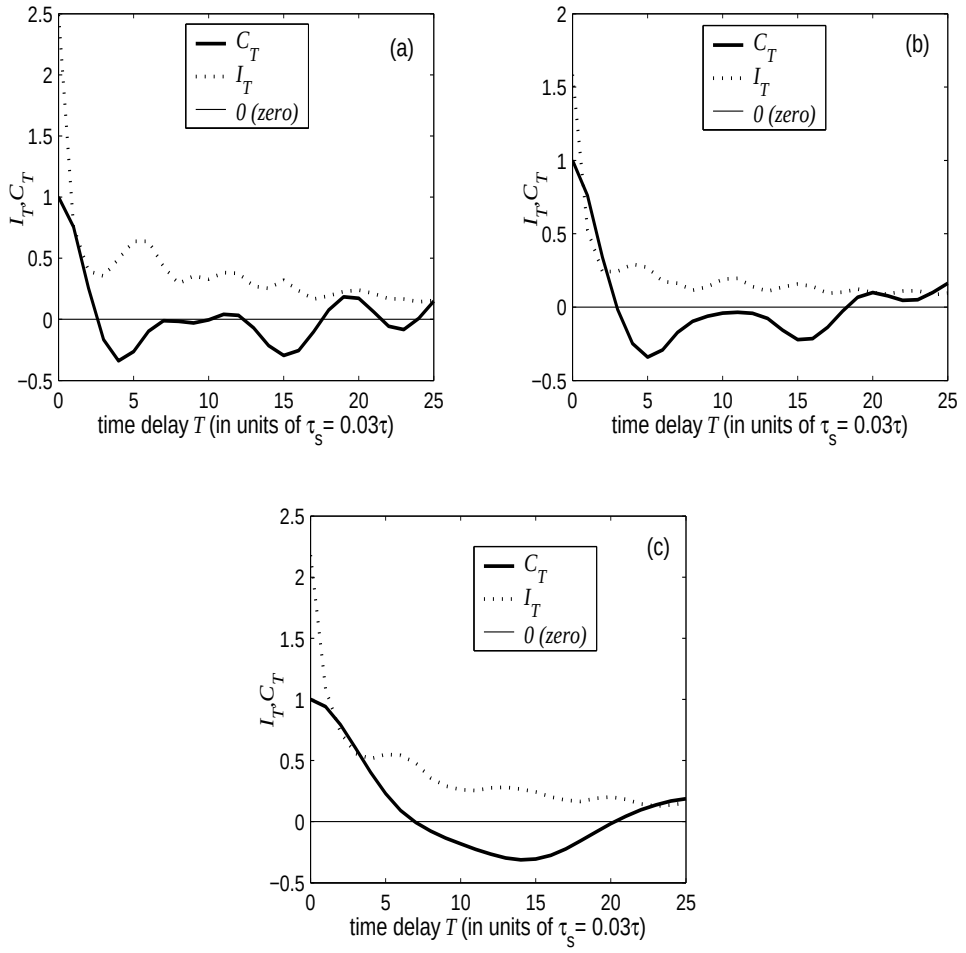


Figure 4.7: Time delay T determination for the coupled CSTR variables (a) $x_j(t)$, (b) $y_j(t)$, and (c) $z_j(t)$ where $j = 1, 2$ is the reactor. The figures include plots for both the autocorrelation function C_T and mutual information I_T.

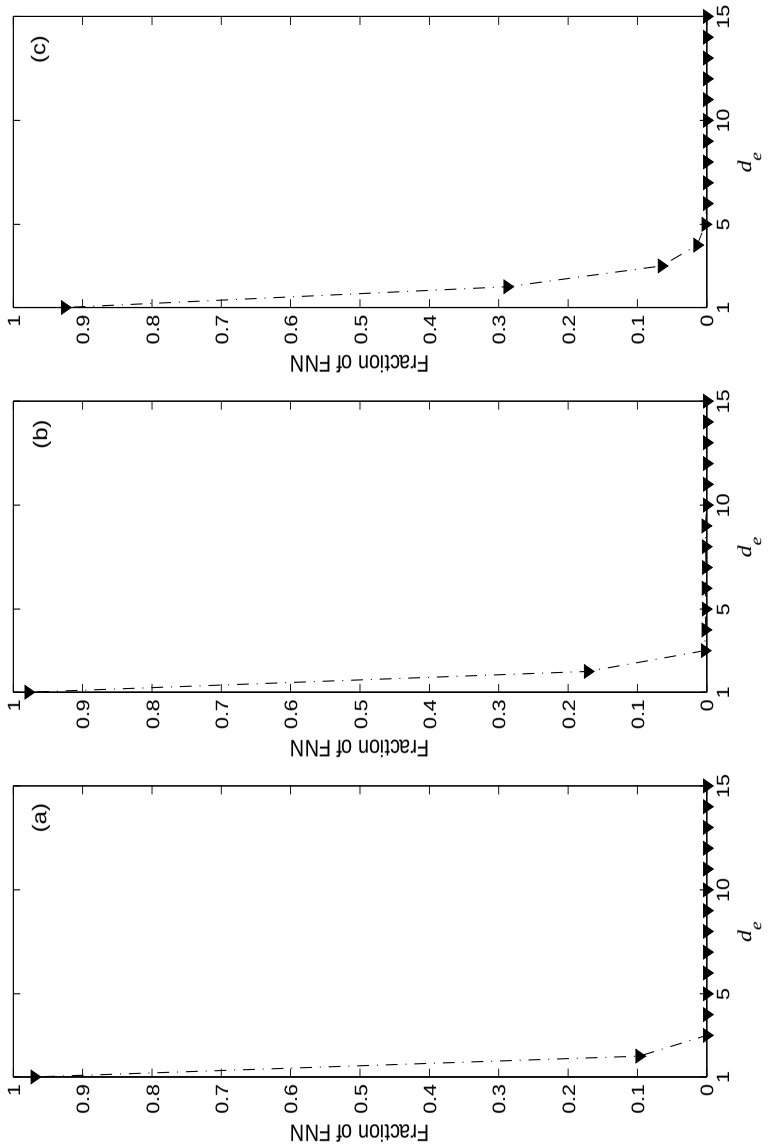


Figure 4.8: Embedding dimension determination using the false nearest neighbours algorithm for the coupled CSTR variables (a) $x_j(t)$, (b) $y_j(t)$, and (c) $z_j(t)$ where $j = 1, 2$ is the reactor.

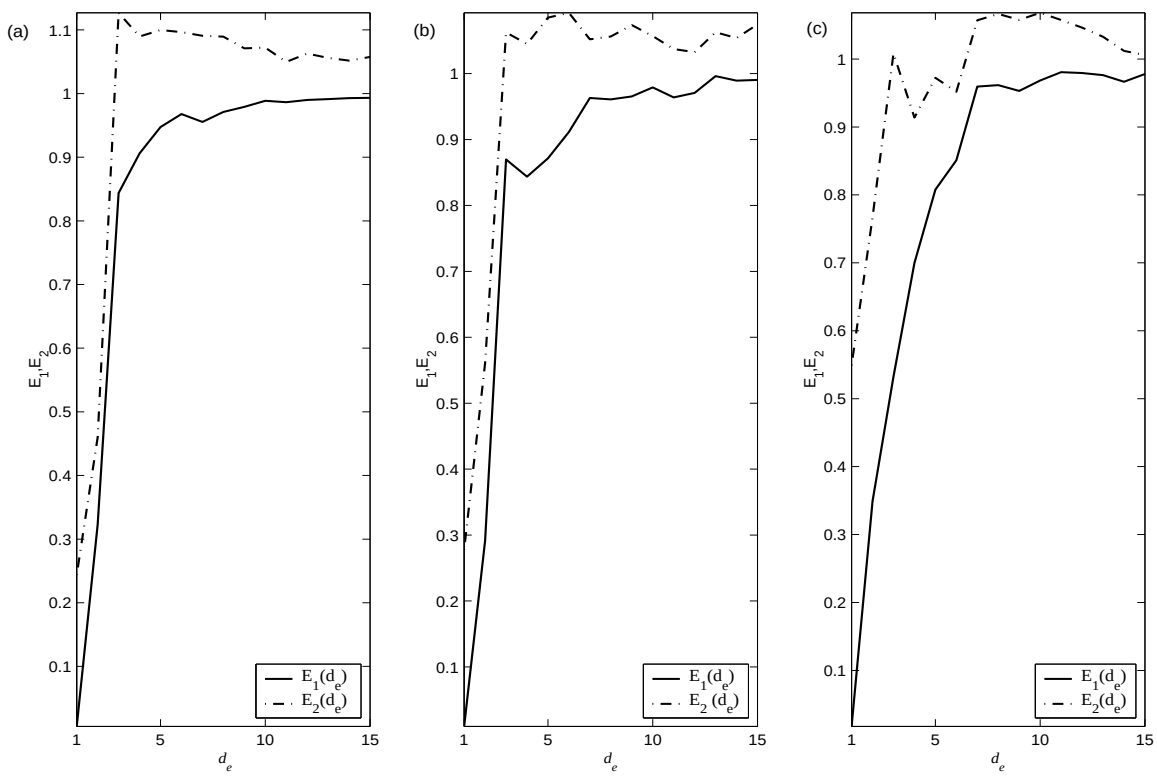


Figure 4.9: Determining the embedding dimension using Cao's method for the coupled CSTR variables (a) $x_j(t)$, (b) $y_j(t)$, and (c) $z_j(t)$ where $j = 1, 2$ is the reactor. E_1 in principle saturates at some d_e. However, for finite and noisy data the saturation point is not easily resolved. The quantity E_2 distinguishes deterministic from stochastic signals. For stochastic signals $E_2 = 1$. For deterministic signals, therefore, there exists some d_e's with $E_2 \neq 0$. Plotting both quantities resolves the minimimum embedding dimension required (Cao, 1997).

4.4 Testing for Nonlinearity

Nonlinear dynamics theory is an interesting and advanced approach that, in principle, enables the classification of chaotic systems. However, the theory and, in particular the "toolkit" for the analysis of time series are still developing. Therefore, there exists a risk of spurious identification of chaos in data that is otherwise consistent with a simple linear explanation. Before use is made of the nonlinear dynamics concepts and tools it is imperative to establish statistical evidence of nonlinearity arising from the underlying dynamics generating the data. *Surrogate data analysis* (Schreiber and Schmitz, 2000; Theiler *et al.*, 1992) is typically used to characterize time series signals generated by nonlinear systems. The method does not identify the existence of nonlinearity directly. Instead, one proves that the data is inconsistent with data generated by certain class of linear systems.

4.4.1 The method of surrogate data

Surrogate analysis is based on the hypothesis testing and bootstrap methods from statistical theory. One proposes a null hypothesis H_0, which is the least interesting explanation that cannot be ruled out from the data. An ensemble of artificial or surrogate data sets consistent with the null hypothesis H_0 is generated. A discriminating test statistic t_{stat} likely to reject H_0 is computed for surrogate data to obtain a range of values associated with 95% ($\alpha = 0.05$)[2] of the test statistic's distribution when the null hypothesis is true. If the value of t_{stat} for the original data falls within the estimated probability distribution of the same statistic calculated for the surrogate data the null hypothesis is accepted; otherwise it is rejected. The basis of rejection is based on calculating a dimensionless quantity, S called the "significance". S is given by the difference between the test statistic for the original data

[2]Other values for the level of significance α may be used depending on the likely error rate of rejecting a true null hypothesis

$\mu_{original}$ and the mean value of the statistic computed from the surrogate data $\bar{\mu}_H$, divided by the standard deviation of the statistic for the surrogates σ_H

$$S \equiv \frac{|\mu_{original} - \mu_H|}{\sigma_H} \tag{4.5}$$

Thus, a significance, say, 2σ's may be not be especially significant whilst higher values are (Theiler *et al.*, 1992).

Different approaches for generating the surrogate data for univariate time series exist and include Fourier-based typical and constrained randomization methods, and a general constrained randomization method based on simulated annealing (Schreiber and Schmitz, 2000). The constrained-randomization method is particularly useful because of certain appealing properties (Theiler and Prichard, 1996). In this scheme, surrogate data *"exactly like"* the original data are generated consistent with a pre-defined null hypothesis H_0. "Exact likeness" is imposed on the surrogate data such that the surrogates mimic all trivial or linear features of the original data. This includes first and second orders moments of the data, auto-correlation function, etc. Possible null hypotheses include (Theiler *et al.*, 1992)

- Temporally uncorrelated (independent identically distributed) noise
- Linearly correlated noise
- Static monotic nonlinear transformation of linearly correlated noise

The null hypotheses of "static monotic nonlinear transformation of linearly correlated noise" is the most general and is used in all results reported in this report.

The method of surrogate data was extended to multivariate times series by Prichard and Theiler (1994). Multivariate surrogate data preserve both the auto-correlations in the individual channels as well as the cross-correlations among the channels. Prichard and Theiler also reported that detection of nonlinearity can

be robust when using multivariate time series than using only each of the individual components. Their **A**mplitude-**A**djusted **F**ourier **T**ransform approach was improved in Schreiber and Schmitz (2000) to an approach where the **AAFT** is repeated in iterative fashion until the differences between the linear statistics are constant at the optimal minimal. A simple calculation showing the steps of generating surrogate data is included in the appendices.

4.4.2 Results of surrogate analysis

The Tisean© package was used to generate surrogate data for all the reported results. The correlation dimension estimate and the 2^{nd}-order entropy were used as discriminating statistics. The correlation dimension was estimated as a function of viewing scale $d_c(\varepsilon_0)$ and as a function of the embedding dimension using Judd's implementation (Judd, 1992) and the Gaussian Kernel Algorithm (GKA) (Yu *et al.*, 2000) respectively. The entropy was also estimated as a function of the embedding dimension using the GKA algorithm. For each embedding, 20 surrogates were generated. Figure 4.10 and Table 4.5 are results obtained using $d_c(\varepsilon_0)$ as the discriminating statistic. Although in all cases investigated the null hypothesis of static monotic nonlinearly transformed linearly filtered noise was rejected, including multivariate information reduced the "significance" of the test. Hence, multivariate surrogate analysis yielded a robust test for nonlinearity.

Table 4.5: Comparison of the significance (S) for nonlinearity testing of univariate and multivariate data surrogates

variable(s)	$\log_e(\varepsilon_0)$	S
x	-1.4967	26.199
xy	-1.4967	24.531
xz	-1.6118	19.439
xyz	-1.6118	16.757

Figures 4.11- 4.13 below are some of the results obtained in the test for nonlinearity using the GKA implementation for computing $d_c(m)$ and $K(m)$, that is

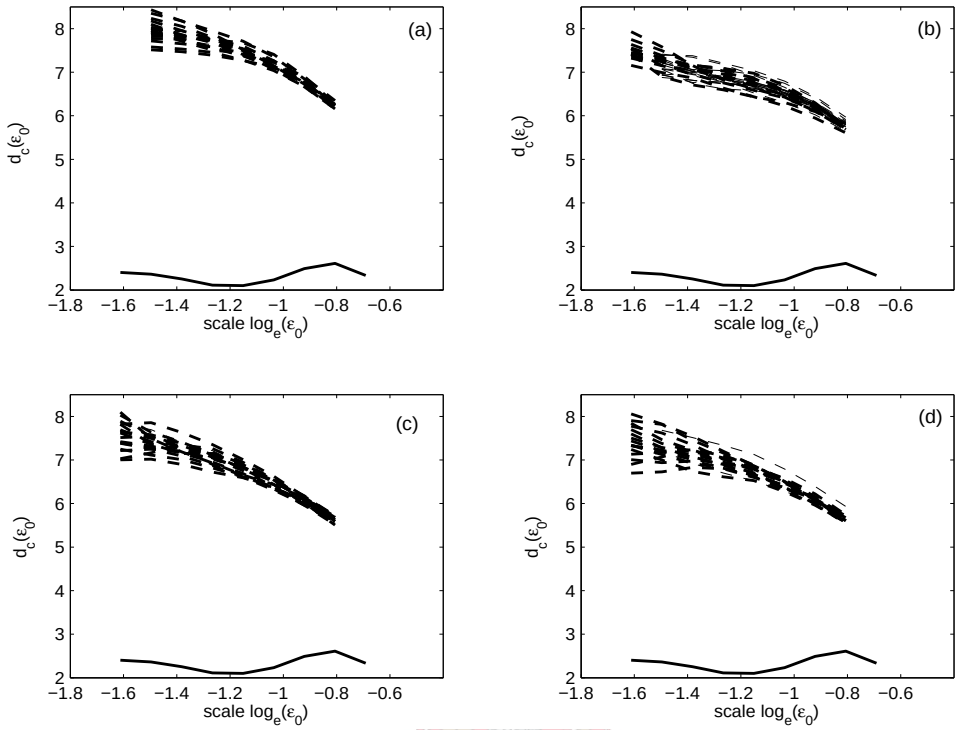


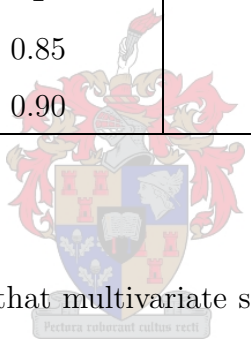
Figure 4.10: Testing for nonlinearity using Judd's Algorithm using the x component for surrogate data (a) ignoring any cross-correlations (b) preserving cross-correlations between x and y (c) preserving cross-correlations between x and z (d) preserving cross-correlations between x, y, and z

the correlation dimension and entropy as a function of the embedding dimension respectively. When cross-correlations were not taken into account the surrogates gave test statistic values with distributions different from the data. Hence, H_0 was rejected. However, preserving cross-correlations in the surrogates resulted in a decrease in the proportion of surrogates with lower values of d_c and correspondingly higher values of K, Table 4.6. As Figures 4.12 and 4.13 show, a few of the surrogate data (3 or 4) gave wildly erroneous K values (< 0). It is, therefore, possible that the surrogate generation algorithms are giving surrogates that are not equally

constrained. Furthermore, the "pivotalness"[3] of the $d_c(m)$ and $K(m)$ appears poor compared to Judd's implementation. It cannot, therefore, be concluded that use of multivariate data improved the robustness of nonlinearity tests when using the GKA algorithm.

Table 4.6: Effect of preserving cross-correlations in nonlinearity testing using the Gaussian Kernel Algorithm for dimension and entropy estimation. $p(\cdot)$ is a proportion

variable(s)	$p(d_c^{surr} < d_c^{data})$	$p(K^{surr} > K^{data})$
x_1	1	1
x_1, y_1	0.85	0.90
x_1, y_1, z_1	0.90	0.85



From the foregoing it is noted that multivariate surrogate generation take into account cross-correlations that may exist between components being analyzed. As shown in Table 4.6 preserving cross-correlations among channels introduced some robustness for the nonlinearity test. However, calculations using the GKA implementation revealed that the surrogate generating algorithms seem not to perform the same amount of randomization when cross-correlations are taken into account. Hence, it is difficult to attribute the robustness of nonlinearity tests using multivariate data to the preservation of cross-correlations because of the unreliability of the surrogates. The issue of the how "pivotal" the test statistic is when using multivariate data also needs to be addressed.

[3] A *pivotal* test statistic yields similar probability density distribution for all processes consistent with the hypothesis.

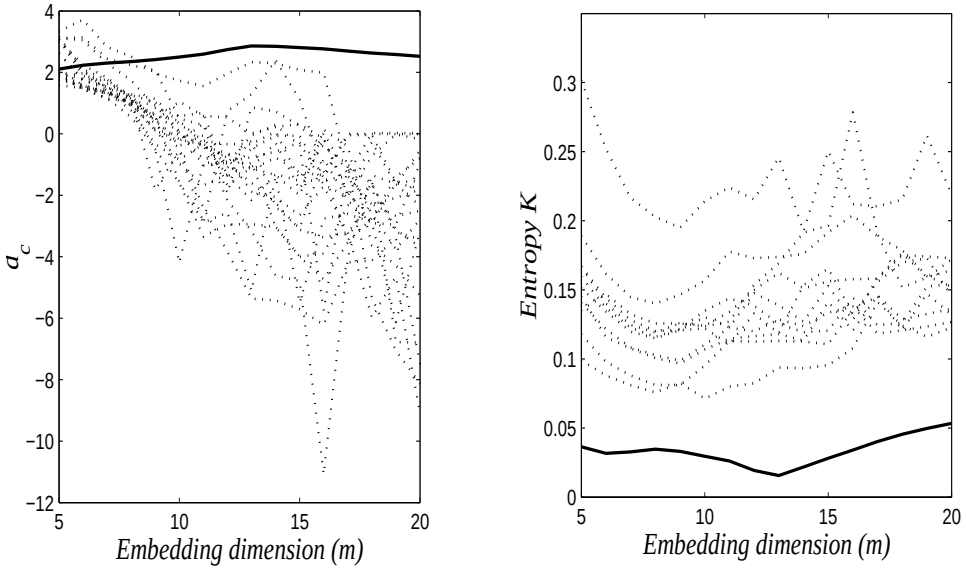


Figure 4.11: Nonlinearity tests using $d_c(m)$ and $K(m)$ as discriminating statistics with x as the observed variable, ignoring any possible cross-correlations with other channels in the of surrogate data

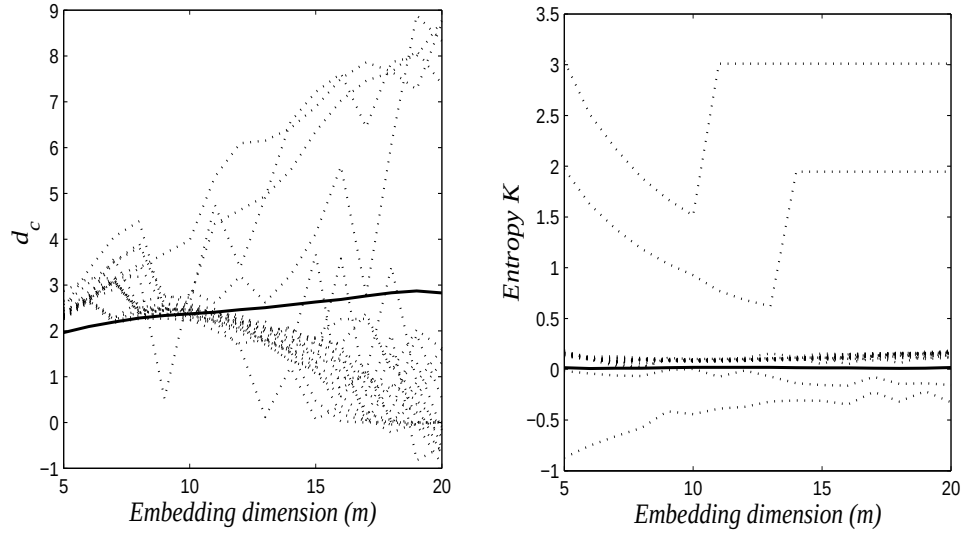


Figure 4.12: Nonlinearity tests using $d_c(m)$ and $K(m)$ as discriminating statistics, with x as the observed variable and preserving cross-correlations between x and y in the surrogate data generation

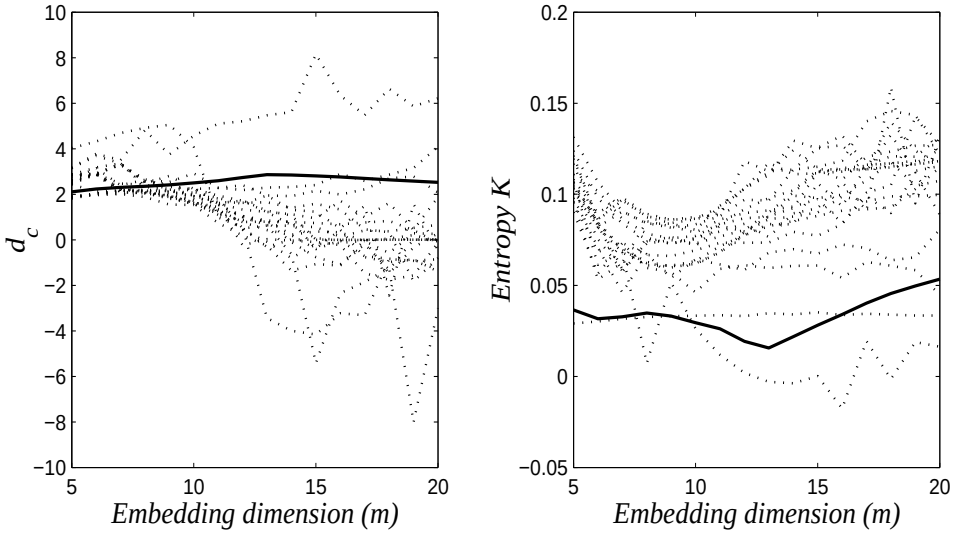


Figure 4.13: Nonlinearity tests using $d_c(m)$ and $K(m)$ as discriminating statistics. Results apply to x as the observed variable and preserving cross-correlations among the x, y, z variables in the surrogate data

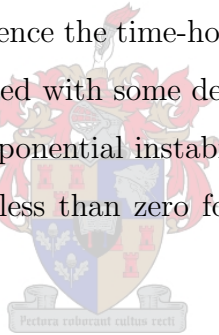
4.5 Estimation of System Invariants

Ergodicity (Eckmann and Ruelle, 1985) guarantees that reconstructing system dynamics preserves topological and geometric invariants of the system. Invariants are important for positive evidence of the possible existence of low-dimensional determinism in physical systems. Invariants that have proved useful in the analysis of physical systems and of which fairly reliable implementations have been developed include ;

- Generalized Dimensions – These are estimates of the number of effective degrees of freedom. Chaotic systems exhibit non-integer dimensions, hence their being termed strange attractors. However, non-chaotic strange attractors with non-integer dimensions also exist. Hence, whilst a chaotic system has a non-integer dimension, the reverse is not necessarily true.

- Entropy (K) – Quantifies the information generation rate, that is, it characterizes the amount of information needed to predict the next measurement to a specified resolution.

- Characteristic or Lyapunov exponents (λ_i) – quantify the sensitivity to initial conditions. Chaotic systems show an exponential growth of infinitesimal changes in the initial condition, simultaneously remaining bounded in a subset of the phase space through a stretching and folding mechanism. To remain bounded, the folding mechanism predominates the stretching mechanism. A dynamical system has as many exponents as the number of coordinates. The spectra of Lyapunov exponents provides information on the sensitivity of a system to perturbations and, hence the time-horizon over which the evolution of the dynamics can be predicted with some degree of confidence. A positive Lyapunov exponent ensures exponential instability. However, the sum of the Lyapunov exponents must be less than zero for boundedness in the case of dissipative systems.



4.5.1 Results and discussion on correlation dimension estimates

The correlation dimension estimate was calculated for different choices of variables in the multicomponent embedding using Judd's implementation. The GKA was also used to confirm the values obtained. The GKA algorithm does not accept multivariate data in its current implementation, hence computations were only done for individual time series. The results of the calculations are summarized in the following figures.

Three distinct patterns can be seen in the correlation dimension estimate plots in Figures 4.14 and 4.16: (a) curves that show a general linear increase of $d_c(\varepsilon_0)$ estimates with decreasing scale ε_0; (b) curves with initial decrease then an increase in $d_c(\varepsilon_0)$ with decreasing viewing scale; and finally (c) curves that initially show

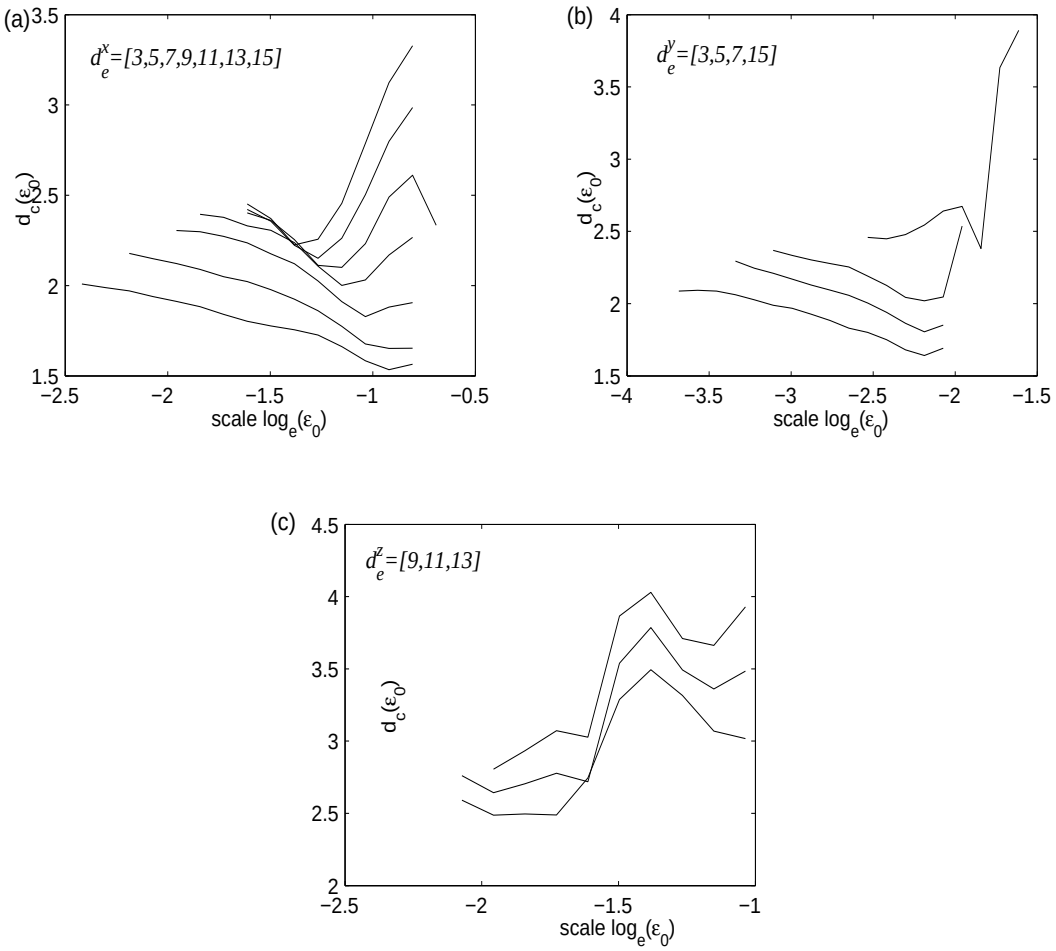


Figure 4.14: Correlation dimension estimates from scalar reconstruction using the variables (a) x (b) y (c) z, using the indicated values for d_e

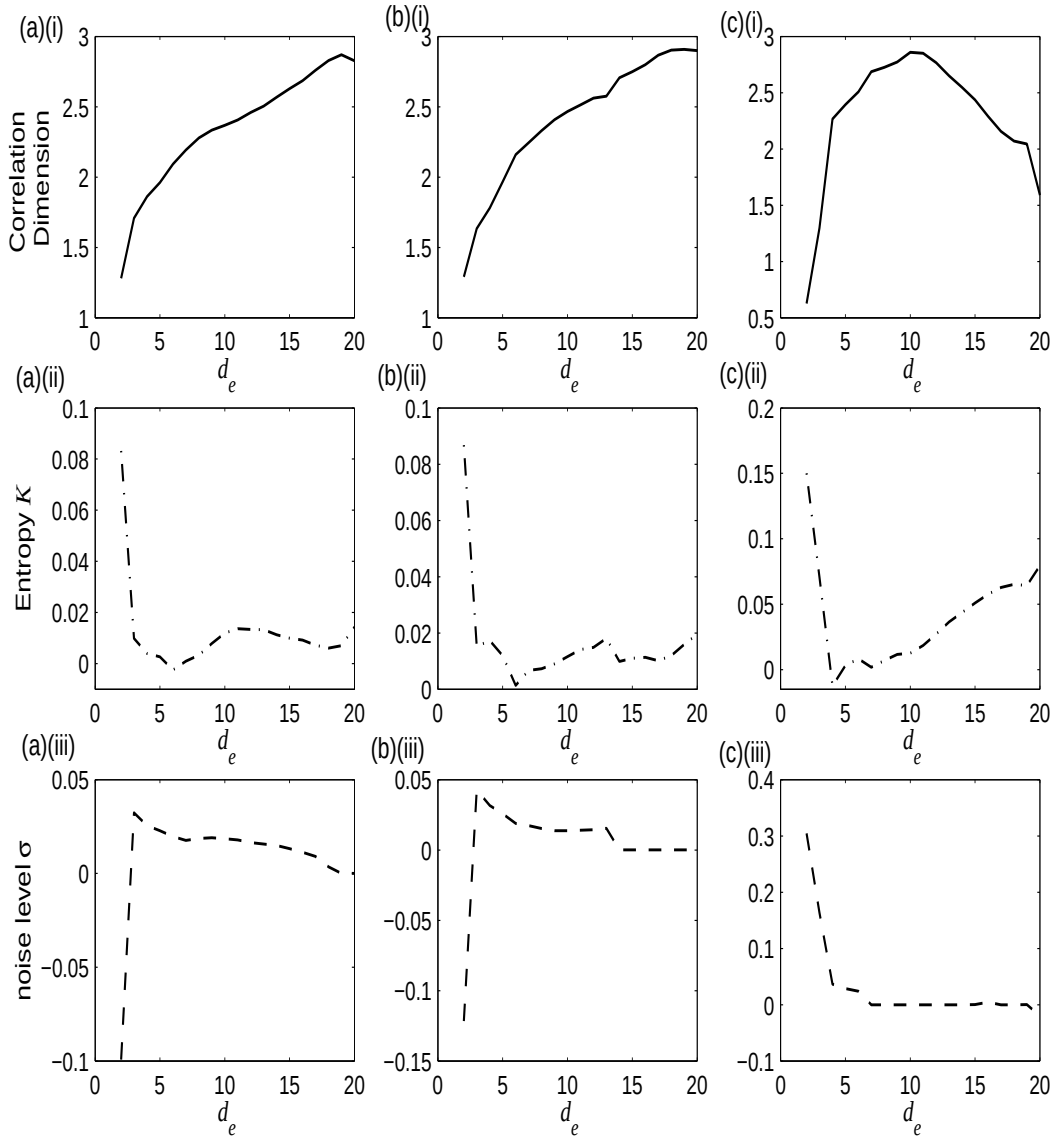


Figure 4.15: Correlation dimension estimates and entropy estimates as a function of the embedding dimension.

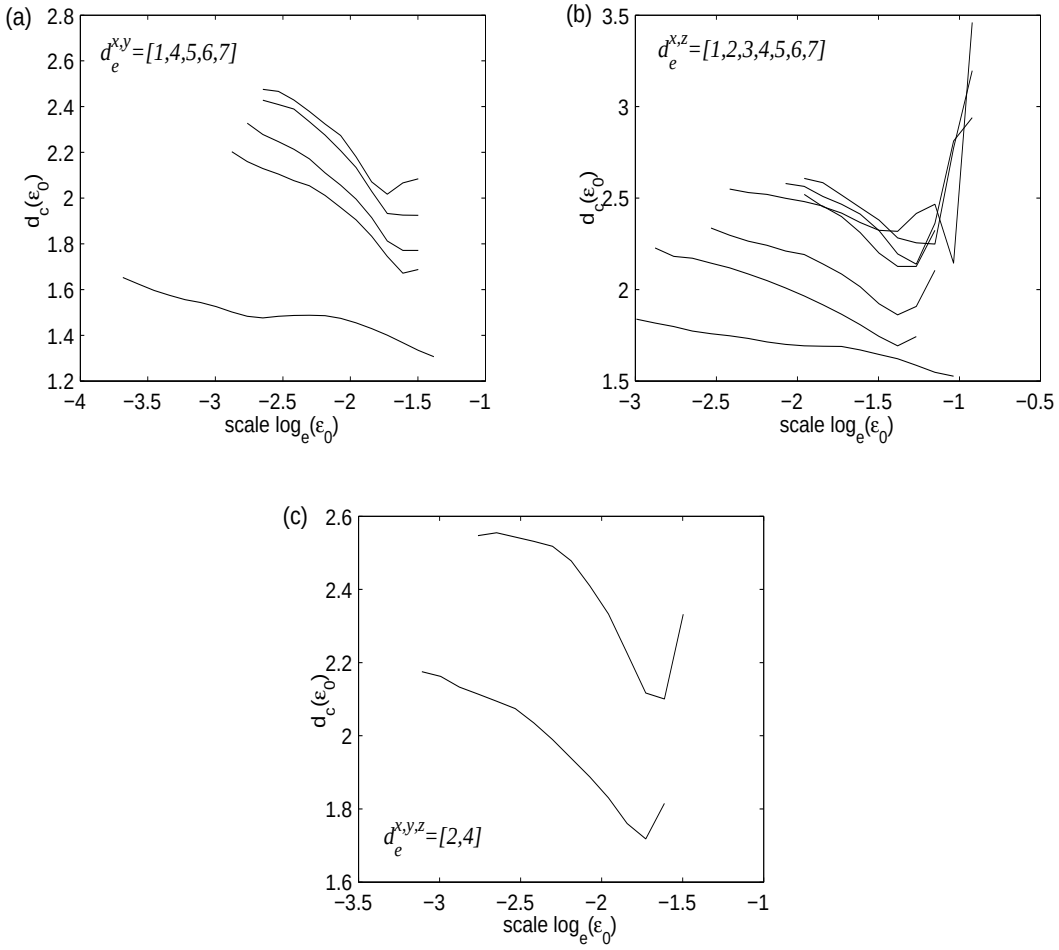


Figure 4.16: Correlation dimension estimates for multivariate embeddings using variables (a) x, and$y$ (b) x and z (c) $x, y,$ and z. In each reconstruction, the same of d_e was used for each component for the values indicated.

fluctuations in $d_c(\varepsilon_0)$ with decreasing scale before stabilizing to a more or less linear or slightly linear increase in the correlation dimension estimates with decreasing scale. Also, all the plots show a general shift to higher dimensions with increasing embedding dimension d_e. The increase is, however, not consistent. Some $d_c(\varepsilon_0)$ values appear to approach a plateau as the embedding dimension increases. At the large scales the dimension estimates are bounded as $1.4 \leq d_c(\varepsilon_0) \leq 4.2$. For lower scales the range is even more restricted, $2 \leq d_c(\varepsilon_0) \leq 3$, suggesting a correlation dimension of about 2.5 for the system. The bounding of the $d_c(\varepsilon_0)$ estimates to values around 2.5 cannot simply be attributed to a attractor volume filling up as d_e dimensions as high as 15 were used yielding similar behaviour as the case for lower embedding dimensions. Furthermore, if such small scale dynamics are due to a random component it would not be expected a tendency of values "clouding" around 2.5. In fact, a plot of the "average" value of d_c with increasing embedding dimension using the gaussian kernel algorithm (which implements Judd's algorithm) shows that at higher embedding dimension the d_c estimates start falling down, which is an artifact of finite data points, Figure 4.15. The GKA algorithm provides an estimate of the noise level in the signal, which is seen to be zero from the figure. Random components show wild fluctuations in $d_c(\varepsilon_0)$ as d_e increases, and the noise level would have indicated increase at high embedding dimensions.

Comparison of the correlation dimension estimates for univariate and multivariate reconstructions, Figures 4.14 and 4.16, did not show significant differences except that, in general, scalar embeddings could not capture dynamics at smaller scales. Note that Judd's algorithm sometimes fails to get values for certain d_e's. Small indicated that this is due to algorithm structural instability and does not convey anything in terms of the fractal structure or dynamical information[4].

[4]Private Communication

4.5.2 Results and discussion on Lyapunov exponents estimates

The Lyapunov exponent quantifies the growth of infinitesimal perturbations. A positive Lyapunov exponent is a positive evidence of the existence of chaos dynamics. Determining the Lyapunov exponents is, however, not easy, particularly for time series data whose differential equations are unknown. Many methods have been proposed in the literature. One approach for estimating the λ_1 is based on the direct use of the available data (Rosenstein *et al.*, 1993, 1994). Figure 4.17 shows a typical plot obtained. Clearly discernable is a sustained general linear increase (excluding the initial linear rise which is due transient effects). Such a linear increase indicates exponential divergence of initially similar points. The maximal Lyapunov exponent λ_1 is obtained from the slope of a straight line fit approximating the general linear increase, normalized by the sampling interval ($\Delta t = 0.03$) between successive data points in the observed signal. Hence, the estimated λ_1 in Figure 4.17 is obtained as;

$$\lambda_1 = \frac{\Delta y}{\Delta x} \approx \frac{2.83}{50 \times 0.03} = 1.88 \tag{4.6}$$

Alternatively, fitting a local linear or global nonlinear model to the embedded data allows for the estimation of the local Jacobian (linearized dynamics) that govern the divergence of initially similar points. Whilst a global nonlinear model is guaranteed to be superior than direct fits of local linear models, obtaining such a model is not always successful. Table 4.7 gives results of the λ_1 and λ_2 obtained from direct local linear fits. At least one positive exponent λ_1 is observed. The reported value is a mean from a number of iterations of the algorithm. Large deviations from the mean were observed in some of the iterations, which made the estimation of an accurate estimate very difficult. Values for the second exponent alternated between positive and negative values although the correct value is known to be positive

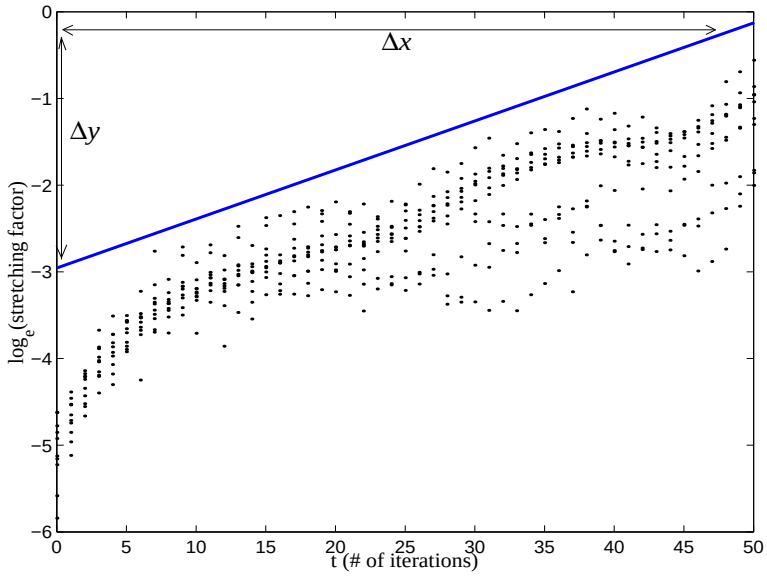


Figure 4.17: Calculating the maximal Lyapunov exponent using Rosenstein algorithm

(Abasher and Judd, 1998). Hence, although knowledge of the maximal exponent is important in identifying the presence of chaotic behaviour in physical systems, estimation of its value using observed measurements is difficult and unreliable.

Table 4.7: Estimates of Lyapunov exponents. Shown are the two largest exponents. A local linear predictor with reference points k was used. D_{KY} is the Kaplan-Yorke dimension, an estimate for the information dimension. Note the close correspondence with $d_c(\varepsilon_0)$ calculated earlier

method	k	λ_1	λ_2
local linear approx. model	30	1.33 ± 0.67	0.00 ± 0.67
local linear approx. model	50	1.16 ± 0.50	-0.83 ± 0.87
local linear approx. model	150	1.16 ± 0.50	-0.50 ± 0.67
D_{KY} estimate $= 2.0 \pm 0.50$			

From the λ_1 values, the system's time evolution can be predicted in time of

about (Abarbanel, 1996);

$$prediction\ horizon \quad = \quad \frac{\tau_s}{\lambda_1} \approx 0.02 - 0.06\tau \tag{4.7}$$

Or, given the sampling time τ_s, there exists, on average, short-term predictability of 3 sampling times ahead. Because of the nature of nonlinear interactions predictability power is lost exponentially fast with an increase in the forecast step ahead. The information dimension estimate given by the D_{KY} has an upper bound that is in agreement with the correlation dimension estimates found earlier. Hence, the validity of the Lyapunov accuracy is thus judged against the correlation dimension estimates and vice versa.

4.6 Information-Theoretic Considerations for Multivariate Time Series

Use of a single time series in inferring system properties is like tossing a coin once — no further information can be extracted without making further *a priori* assumptions. Simultaneously measuring several observables with (possibly) different physical meanings could simplify the analysis if the variables span the original state space. However, measured variables may be complicated functions of the underlying state variables. In this case, the reconstructed attractor is bound to be more folded than in the delay embedding of one coordinate. Information theoretical statistics, such as the Shannon and Kolmogorov entropies, are widely used in nonlinear time series analysis. Statistical dependence between signals (or time delayed copies of the same signal) is often quantified by their mutual information. Schreiber (2000) introduced a relative entropy concept to analyze dynamical properties, such as driving and responding variables, that is based on the consideration of transition probabilities. A simple application of this concept is in deciding the gain of information

when one uses either a simple embedding of only time delayed copies of a single time series or multivariate embedding.

For example, in the case of a single reactor in the coupled CSTR system and in the absence of future information, if one is trying to predict $x(t+\tau_p)$, the relative entropy identifies which of the variables $x(t+\tau)$, $y(t+\tau)$, or $z(t+\tau)$ yields more information in the reconstruction.

$$I^{abb}(\tau, t_p) = \sum_{i_a, j_b, k_b} p(i_a, j_b, k_b) \ln \frac{p(i_a, j_b, k_b)}{p(i_a, j_b) p(k_b)} \tag{4.8}$$

$$I^{bb}(t_p) = \sum_{i_a, j_b} p(i_a, j_b) \ln \frac{p(i_a, j_b)}{p(i_a) p(j_b)} \tag{4.9}$$

where I^{abb} and I^{bb} are the three-point and two-point average mutual information quantities and $p(\cdot, \ldots, \cdot)$ are the the respective joint probability estimates.

Figure 4.18 shows that knowing either y_t or z_t in addition to x_t gives considerably more information about the future x_{t+p} than when only x_t is known. Also, the gain of information from z is much higher than for x. This is evident from the ordinary differential equation describing the system, equation (4.4), where there is a direct relationship between x and z, and none between x and y except through z. However, making use of past values of x_t captures more information compared to the other variables. Intuitively, making use of both time delay information and other variables reduces uncertainty in predicted values.

4.7 Fitting Nonlinear Models to Observed Data

An understanding of the underlying physico-chemical principles makes it possible to develop fundamental models of physical systems. However, such a fundamental approach is difficult, and is often inaccurate and non-robust.Therefore, *empirical*

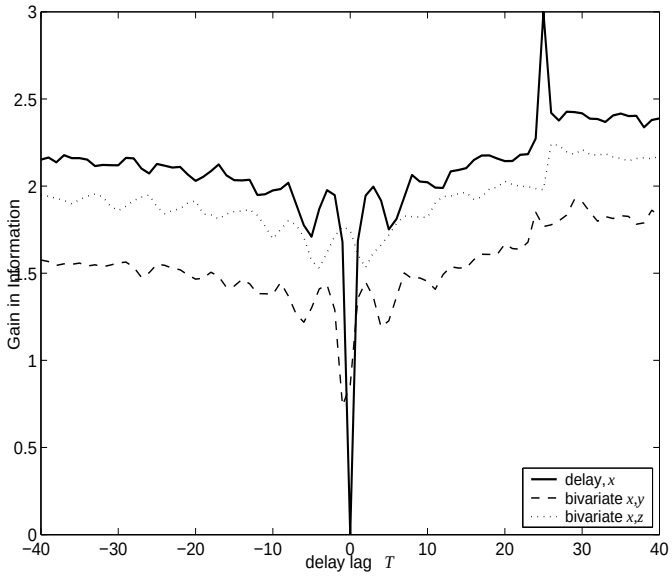


Figure 4.18: Gain of information about the future of x, 25 time steps ahead given $x(t_0)$ at the present time and additionally $x(t)$ (solid) or $y(t)$ (dash dot) or $z(t)$ (dotted) at any other time.

models developed using measurements taken from the system are used to give an insight in the governing dynamics. It should be emphasized that these models are complementary and do not substitute for fundamental models derived from first physical principles. In the next section is an outline the procedure used in approximating predictive models for the time evolution of the x_1 variable of the coupled CSTR system. Figure 4.19 is a summary of the model fitting procedure. The simulated signals $x_j, y_j, z_j, j = \{1, 2\}$ were taken as the observed time series.

4.7.1 Modeling Procedure

I. Using the time series data, the phase space was reconstructed with the parameters as determined in Tables 4.3 and 4.4 for uniform and non-uniform embedding strategies respectively.

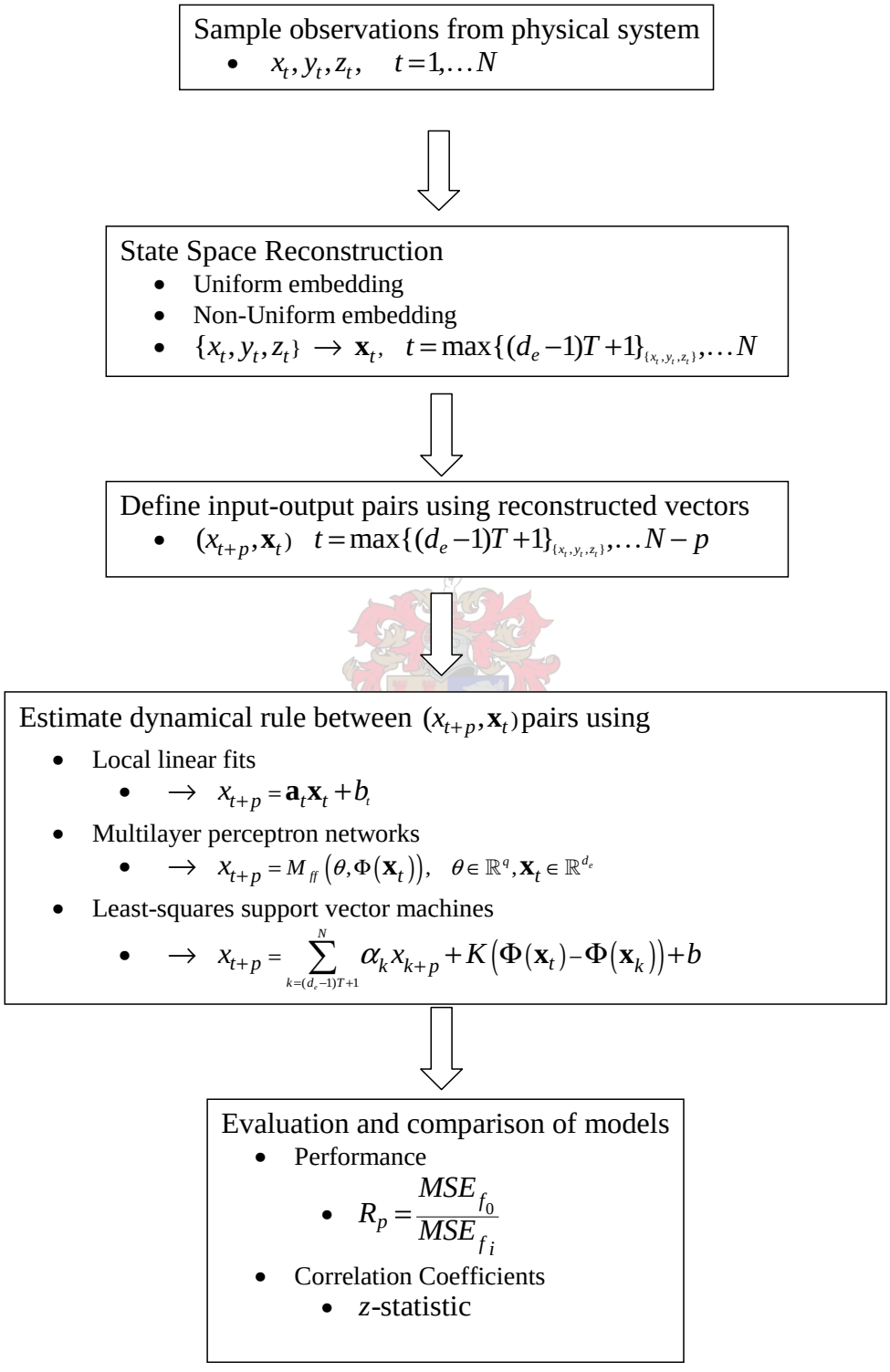


Figure 4.19: Outline summary of the nonlinear model fitting procedure.

II. Learning or training pairs $(x_{t+p}, \mathbf{x}_t)$ were defined such that

$\mathbf{x}_t$ – embedded state space vector

x_{t+p}– the value of x located p time steps ahead of the latest component in $\mathbf{x}_t$.

III. An optimal nonlinear function f defining a dynamical relationship between x_{t+p} and $\mathbf{x}_t$ was fitted according to

$$x_{t+p} = f(\Phi(\mathbf{x}_t)) \tag{4.10}$$

where p, f, Φ, are respectively the prediction time step (usually $p = 1$), nonlinear function approximator, and subspace dimension reduction (PCA or ICA). In the absence of subspace dimensionality reduction $\Phi = 1$. Three different model classes were used in estimating the function f, that is, local linear fits, MLPs, and LSSVMs. For each model class the optimal f was found by tuning various attributes associated with the class as explained in the following.

- **Local linear fitting**

  Local linear approximation methods assumes the relation in equation (4.10) is, to a good approximation, described by an unknown smooth function f. Thus, instead of finding this function improved predictions can be obtained by fitting local linear models of the form

  $$\hat{x}_{n+1} = \mathbf{a}_n \mathbf{x}_n + b_n \tag{4.11}$$

  where $\mathbf{a}_n, b_n$ are parameters for each $\mathbf{x}_n$. This is achieved by minimizing the *non-regularized* least squares function

  $$e^2 = \sum_{\mathbf{x}_i \in \Omega_n} (x_{i+1} - \mathbf{a}_n \mathbf{x}_i - b_n)^2 \tag{4.12}$$

  with respect to $\mathbf{a}_n, b_n$, where Ω_n is some ϵ-neighbourhood of $\mathbf{x}_n$, excluding $\mathbf{x}_n$.

- **Multilayer Perceptron (MLP) Neural Networks**

  A feedforward neural network with a single hidden layer with number of nodes chosen in the range 2 and 30 inclusively selected according to the MSE criterion was used to define a nonlinear function mapping the vector input to a single output, p time steps ahead of the latest component in the vector input. The equivalent of equation 4.10 using MLP is given by

$$x_{t+p} = M_{ff}(\theta, \Phi(\mathbf{x}_t)), \qquad \theta \subset \mathbb{R}^q, \mathbf{x}_t \subset \mathbb{R}^{d_e} \tag{4.13}$$

  where $M_{ff}(\theta, \Phi(\mathbf{x}_t))$ is the parameterized MLP model, θ the synaptic and bias weights parameter vector, and q the model order. The optimal model order q is usually chosen such that the model complexity is minimized. Information theory criteria such as the Bayesian Information Criterion (BIC) and Rissanen's Minimum Description Length principle (MDL) are commonly used to find the optimal order. The order of the MLP model was not optimized to allow for fair comparison of the performance of MLPs against the LSSVMs since comparable BIC or MDL criteria for LSSVMs have yet to be developed. Table 4.8 shows some of the MLP settings used. The training, validation, and test sets were selected as outlined in the section on LSSVM fitting below.

- **Least-Squares Support Vector Machines (LSSVMs)**

  Given the learning pairs $(x_{t+1}, \mathbf{x}_t)$, the LSSVM function fitting is formulated as

$$x_{t+p} = \sum_{k=1}^{N} \alpha_k x_{k+p} K(\mathbf{x}_t - \mathbf{x}_k) + b \tag{4.14}$$

  where K is the kernel function, for which the Gaussian radial basis function was used, see Table 3.1. The regularization (γ) and kernel parameters (σ) were optimized with respect to the mean square error criterion.

Table 4.8: MLP architecture attribute settings

Attribute	Setting
Number of hidden layers	1
Number of output targets	1
Hidden layer transfer function	logistic sigmoid $\left(\frac{1}{1+e^{-x}}\right)$
Output layer transfer function	linear
Training function	Levenburg-Marquadt
Performance function	MSE
Training performance goal	$1.00E^{-06}$

Selection of the best choices for the (γ, σ) pair values is a delicate matter and requires enormous computer processing time with long training data. To avoid the computational cost, the hyperparameters were optimized over a representative subset of the training data. The overall LSSVM model fitting procedure followed the sequence;

i. The observed data was split into three sets: the first 50% of the time series for the training data, the next 25% for the validation test set and the last 25% for the test data. In this case, the lengths of the training, validation, and test sets were 2000, 1000 and 1000 points respectively. (Similarly for the MLP case).

ii. To determine the hyperparameters, a representative subset of the embedded training data with about 500 points was selected by sampling every fourth input-output pair.

iii. Initial hyperparameter candidate tuning sets were defined as $\Gamma = [0,\ 1,\ 5,\ 10,\ 50,\ 100,\ 1000]$ for the regularization constant, and $\Sigma = [0.5,\ 1,\ 2,\ 4,\ 8,\ 16,\ 32,\ 64]$ for the kernel width.

iv. For each possible (σ, γ) pair 10-fold cross-validation was performed. An initial optimal pair (σ_0, γ_0) that gave the least mean square error was selected.

v. Using (σ_0, γ_0) from step (iv), the parameters were further optimized by defining a locally refined grid around the initial optimal parameters to obtain the final optimal pair (σ_1, γ_1).

vi. An LSSVM model was then fitted using the entire training data and the selected hyperparameters (σ_1, γ_1) from (v).

vii. Finally, the performance of LSSVM model was evaluated using an independent test set not used in the training or validation of the fitted model.

The least squares support vector solution is optimal in the case of an approximately normal distributed error function, that is, $e_k \approx N(0, \sigma^2)$. This follows from the *maximum likelihood estimation* principle from statistical theory. An improvement on the LS–SVM model can be obtained by weighting the errors e_k for each support vector such that the resultant error distribution is approximately normally distributed (Suykens *et al.*, 2000). This can be easily done by noticing that $\alpha_k = \gamma e_k$, equation (3.30). Weighting the support vectors results in a γ_k for each of the support vectors k. A histogram fit of the α_k/γ values did not show significant deviation from a normal distribution, Figure 4.21(e). Furthermore, sample weighted models performed as well as the unweighted models with respect to the mean square error. Therefore, it was not considered necessary to obtain robust models through weighting of the errors.

IV. The relative performance of the MLP and LSSVM models were compared

using a *performance index* R_p defined by

$$R_p = \left(\frac{MSE_{f_0}}{MSE_{f_i}} \right) \tag{4.15}$$

where MSE_f is the mean square error of the *base* model used to compare against other models and MSE_{f_i} is the mean square error of model i under consideration. Since the goal was to compare the relative advantages of multivariate embedding and scalar embedding, the base model f_0 was chosen to be that defined on scalar embedding of the x_1 variable using the reconstruction parameters determined previously, that is, $d_e = 11, T = 3$. Thus, using the concept of a R_p and depending on its values the following conclusions could be made;

$$R_p \begin{cases} < 1, & f_i \text{ is a worse predictive model than } f_0, \\ = 1, & f_i \text{ is as good a predictive model as } f_0, \\ > 1, & f_i \text{ is a better predictive model than } f_0. \end{cases} \tag{4.16}$$

V. Finally, the significance test for the differences in the correlation coefficients of MLP and LSSVM models for the same reconstruction strategy were computed. Concepts from the sampling theory of correlation were used to define a test statistic z based on Fisher's Z transformation (see appendices for details). Such a test statistic allows one to formulate hypotheses tests for specified significance levels, which are used to either accept or reject a null hypotheses H_0.

4.7.2 Modeling results

(a) ***Local linear modeling***

Figure 4.20 is a plot of the variation of the relative forecast with increasing neighbourhood size using local linear modeling[5] fit on the different embedding

[5]C code implementation used courtesy of Rainer Hegger

strategies as indicated. The embedding parameters from Tables 4.3 and 4.4 were used in corresponding reconstructions. The raw data was used "as is", i.e., with trivial embedding with $(d_e, T) = (1, 1)$ per each of the six channels $x_j, y_j, z_j,\ j = \{1, 2\}$.

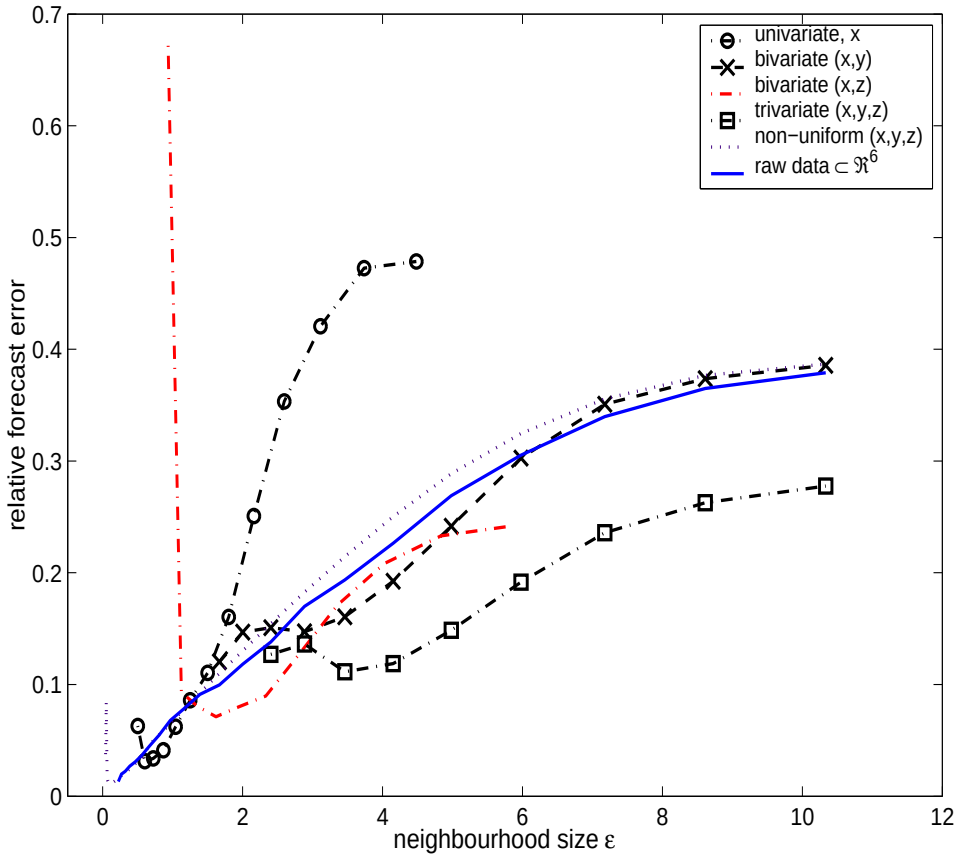


Figure 4.20: Local linear modeling: Variation of relative forecast error with neighbourhood size for different embedding strategies as indicated

(b) ***Nonlinear model fitting*** Summary of results obtained from MLP and LS–SVM modeling are presented in Tables 4.9, 4.10, 4.11 and 4.12. Figure 4.21(a)–(f) are plots obtained from typical LS–SVM and MLP runs, in this case for a model based on (x, y)-bivariate embedding.

[6]PCA – principal component analysis, W – whitening, and ICA – independent component analysis

Table 4.9: Global nonlinear modeling results of the coupled CSTR using MLPs. Also shown are the significance test of correlation for the hypotheses test $H_0 : \rho_i \leq \rho_0, H_1 : \rho_i > \rho_0$, where ρ_0 is the correlation coefficient for base model f_0

variables(s)	d_e	T	R_p	$\hat{\rho}$	Z_i	z
x_1 (base model f_0)	11	3	1.00	0.9972	3.287	0.00
x_1, y_1	[11 13]	(3,2)	8.25	0.9999	4.952	36.59
x_1, z_1	[11 11]	(3,4)	0.87	0.9972	3.285	0.04
x_1, y_1, z_1	[11 13 11]	(3,2,4)	11.66	0.9998	4.557	27.85
$x_1, y_1, z_1, x_2, y_2, z_2$	[1 1 1 1 1 1]	(1,1,1,1,1,1)	107.98	1.0000	5.554	50.20

Table 4.10: As in Table 4.9, but using LSSVM with Gaussian radial basis kernels

variables(s)	d_e	T	R_p	$\hat{\rho}$	Z_i	z
x_1 (base model f_0)	11	3	1.00	0.9982	3.509	0.00
x_1, y_1	[11 13]	(3,2)	2.35	0.9992	3.931	9.28
x_1, z_1	[11 11]	(3,4)	1.19	0.9985	3.601	2.01
x_1, y_1, z_1	[11 13 11]	(3,2,4)	1.85	0.9990	3.815	6.72
$x_1, y_1, z_1, x_2, y_2, z_2$	[1 1 1 1 1 1]	(1,1,1,1,1,1)	4.02	0.9996	4.200	15.30

Table 4.11: MLP and LSSVM modeling results using a non-uniform embedding approach. The z statistic for determining the significance in correlation differences uses reference models f_0 in Tables 4.9 and 4.10.

Embedding			MLP			LSSVM		
variable(s)	lag vector	d_e	R_p	$\hat{\rho}$	z	R_p	$\hat{\rho}$	z
x_1	l_x	12	5.53	0.9995	19.84	6.49	0.9997	72.73
x_1, y_1	l_y	16	63.66	1.0000	46.7931	7.98	0.9988	75.03
x_1, z_1	l_z	8	13.30	0.9998	26.69	54.11	1.0000	97.70
x_1, y_1, z_1	l_z	12	96.08	1.0000	50.08	86.70	1.0000	102.17

Table 4.12: Comparison of dimension reduction methods on the basis of one-step MSE of the fitted model and significance of the differences in the correlation coefficients under the hypotheses test $H_0: \rho_i = \rho_0, H_1: \rho_i \neq \rho_0$.

Dimensionality Reduction		MLP			LSSVM		
Projection	Method[6]	R_p	$\hat{\rho}$	z	R_p	$\hat{\rho}$	z
$\mathbb{R}^6 \rightarrow \mathbb{R}^4$	PCA	1.00	0.9685	0.00	1.00	0.9705	0.00
$(x_1, y_1, z_1, x_2, y_2, z_2)$	PCA+W	0.93	0.9728	1.69	1.06	0.9699	0.07
	ICA	0.92	0.9361	8.08	1.06	0.9717	0.64
$\mathbb{R}^{11} \rightarrow \mathbb{R}^{11}$	PCA	1.00	0.9974	0.00	1.00	0.9981	0.00
(x)	PCA+W	0.6863	0.9976	0.97	0.9882	0.9981	0.11
	ICA	0.9722	0.9984	5.32	0.9662	0.9980	0.40
$\mathbb{R}^{24} \rightarrow \mathbb{R}^{12}$	PCA	1.00	0.9990	0.00	1.00	0.9973	0.00
(x, y)	PCA+W	0.75	0.9987	3.22	0.95	0.9972	0.56
	ICA	0.71	0.9883	27.49	0.79	0.9966	2.60
$\mathbb{R}^{35} \rightarrow \mathbb{R}^{15}$	PCA	1.00	0.9987	0.00	1.00	0.9976	0.00
(x, y, z)	PCA+W	0.65	0.9987	0.76	1.08	0.9978	0.9
	ICA	0.61	0.9992	4.84	1.09	0.9978	1.05

4.7.3 Discussion

(a) ***Local linear modeling***

Figure 4.20 showing the variation of relative forecast error with neighbourhood size has optima at small ϵ-sizes for all plots. Casdagli (1991) suggested a test for nonlinearity based on the variation of the relative forecast error with neighbourhood size on which the fit in equation (4.11) is made. If the optimum occurs at large neighbourhood then the embedded data are best described by a linear stochastic process. Correspondingly, the occurrence of the optimum at smaller neighbourhood sizes indicated a nonlinear deterministic process.

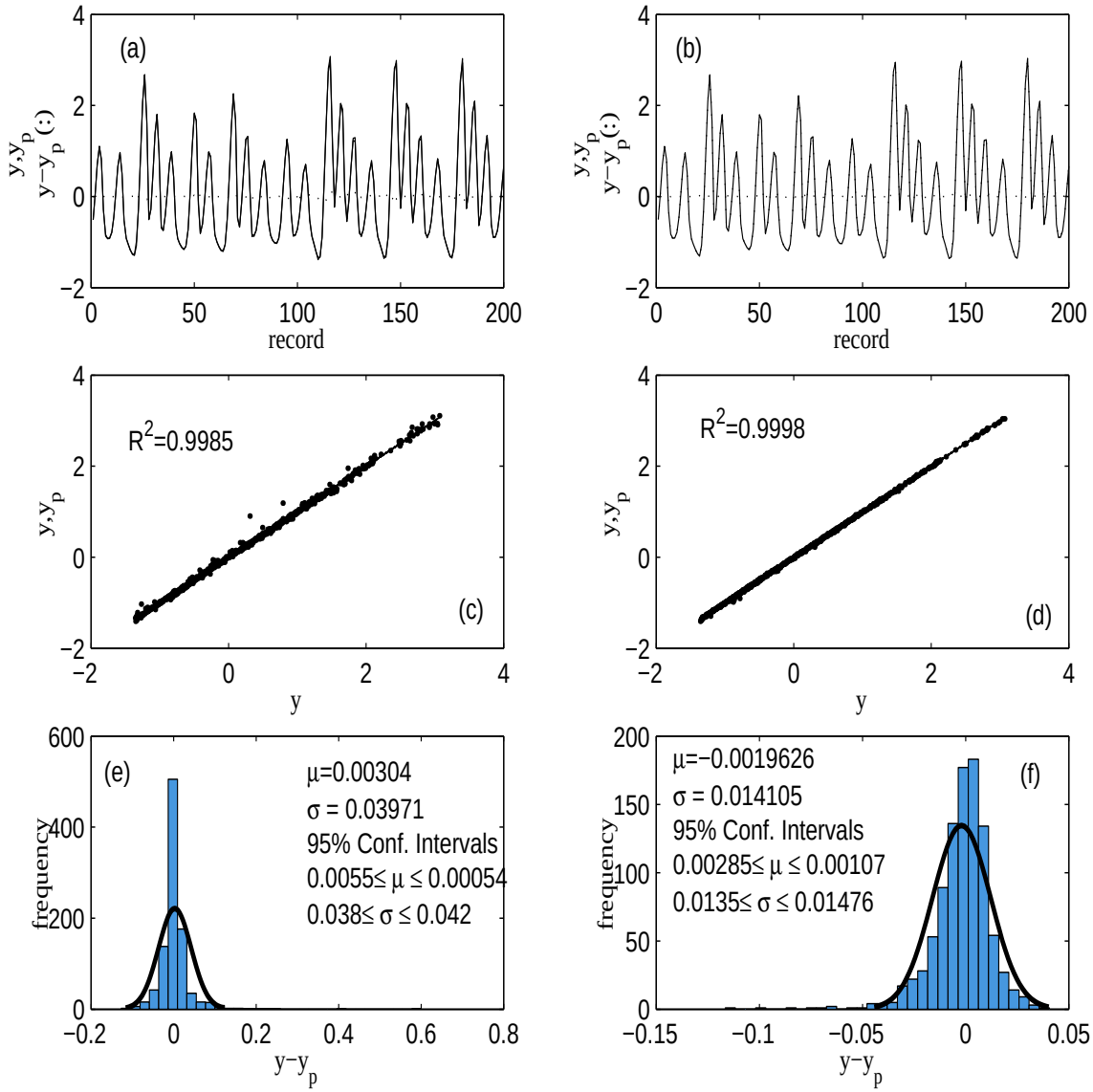


Figure 4.21: Typical LS–SVM and MLP modeling results. (a) and (b) plots of predicted and actual output for LSSVM and MLP models respectively. (c) and (d) are corresponding regression plots between the actual data and model output. (e) and (f) Probability distribution plots for prediction errors for respective models. A normal p.d.f. fit is shown superimposed on the plots.

Here, the concept was extended to local models for multivariate embeddings. It was observed that the data (in the respective embedding spaces) were consistent with a nonlinear deterministic generator. The univariate embedding strategy was unable to retain a deterministic structure at large neighbourhood sizes. In contrast, relatively strong determinism in multivariate embeddings was evident in large ϵ- neighbourhood sizes. Acknowledging the fact that most approaches in nonlinear dynamical analysis exploit the information in some neighbourhood, the results indicated that inclusion of simultaneously measured data resulted in better capture of the system dynamics over a broader ϵ-spectrum compared to scalar embeddings.

The rate of increase of the relative forecast error for the univariate x-embedding was largest, increasing exponentially with neighbourhood size. Beyond $\epsilon \approx 1.5$ univariate embedding had the worst performance. (x, y, z)-trivariate non-uniform embedding (dotted line) and true state space embedding (solid line) exhibited the same behaviour, overlapping in small neighbourhood sizes. Moreover, both had minima at the smallest neighbourhood size and, therefore, the lowest relative forecast error over all the embedding strategies considered. This indicates that non-uniform embedding strategy capture better the dynamics at infinitesimal scales. The error growth rates of the multivariate embeddings with increasing ϵ were more or less similar. The (x, z)-bivariate embedding, however, is less certain. It exhibits fluctuations, which can be attributed to possible lack of enough data points in the embedded space.

Retention of dynamics at larger scales is not as important as at smaller scales due to the effect of sensitivity to initial conditions. In spite of this, local linear modeling results indicate a hint that reconstruction methods that "minimize" the stretching effect inherent in the dynamics make better information generators, i.e., the trajectory paths diverge less fast than for uniform embeddings.

Hence, multivariate embedding approaches capture both short and long-range effects to yield better local linear fits. Nonetheless, it must be noted that for all neighbourhood sizes indicated reconstruction were superior to using the data average for the one-step ahead forecast (whose relative forecast error $= 1$, for all ϵ).

Therefore, with respect to local linear modeling and in the absence of sufficient data, multivariate embedding were superior to univariate. This is due to their capacity of retaining information of the deterministic structure in both small and large neighbourhood sizes. Moreover, non-uniform multivariate embedding yielded fits that traced almost perfectly the behaviour of the true underlying trajectory in smaller neighbourhood sizes.

(b) ***Global non-linear modeling***

Tables 4.9 and 4.10 show a generally better performance of multivariate models over the univariate case. In the case of MLP models, inclusion of simultaneously measured information from the y_1 variable resulted in an 8-fold decrease in the mean square error. A phenomenal improvement in the performance was observed in the case where all six (6) original variables were used to train the network. However, inclusion of information in the z_1 variable for predicting the evolution of the $x_1(t)$ variable had a negative effect on the performance. Statistical analysis of the correlation coefficients further confirm this. Under the null hypothesis $H_0 : \rho_{f_i} \leq \rho_{f_0}, H_i : \rho_{f_i} > \rho_{f_0}$, values of the z statistic as calculated were greater than 2.33 (the corresponding value for a one sided test at a significance of 0.01) in almost all cases except for the (x, z)-embedding. Hence, the null hypotheses that the correlation coefficients are equal to or less than for the univariate embedding was rejected in all but this single case.

The same pattern was evident in LSSVM results though not as dramatic as in the MLP case, Table 4.10. Additionally, whilst the inclusion of variable z did not result in significant improvement in performance compared to the y variable, it had a marginal positive effect. The test for significance in the difference of the correlation show that the null hypothesis that $\rho_{f_0} = \rho_{f_{xz}}$, where $\rho_{f_{xz}}$ is the correlation of the multivariate model using the x, y variables, could be rejected at the 99% confidence level. For the other cases the null hypothesis was rejected.

Non-uniform embedding strategies showed an improvement in the mean square error and correlation statistics compared to the corresponding base model f_0 for both MLPs and LSSVMs, Table 4.11. This exceptional performance is a direct consequence of the fact that multiple time scales, otherwise "unseen" by uniform embedding, are taken into account. The optimal delay (T) selection methods for uniform embedding identify a single time scale (*cf.* the dominant frequency f_1 obtained in Fourier Analysis). Moreover, methods for determining a suitable (d_e, T) pair are generalization of heuristic techniques proven optimal under certain specific contexts, e.g., T corresponding to the first zero of the mutual information is optimal only for 2-dimensional embeddiings. Non-uniform embedding is based on the idea that the reconstruction process cannot be optimized in isolation from the intended use of the embedded vectors. Hence, embedding should be optimized concurrently with the modeling objective. The approach taken here only partially resolved this issue. Further improvements will be to optimize the selection procedure within the nonlinear modeling process itself. This has been done successfully for radial basis function networks, where it was shown that variable embedding yield cylindrical basis functions that trace the trajectory better (Judd and Mees, 1998; Small, 1998). Further theoretical work to extend the concept to

LSSVMs and MLPs is required. A variation of the variable embedding strategy is expected given the computational cost incurred in using MLPs and especially LSSVMs. The conclusions drawn are certainly true with respect to the specific cases of global nonlinear models used here but other model classes are expected to yield similar results.

Performance of one-step ahead prediction is not indicative of the long-term behaviour of the dynamics. Free-run prediction is a more powerful test for evaluating the performance of a model under iteration(Barnard, 1999). For multivariate nonlinear time series analysis, this requires building either models that predict as many variables as the observed variables simultaneously or, alternatively, constructing predictor models for each observed variable. Given the nature of chaotic or nonlinear systems a difficulty arises in that one has to ensure such vector prediction models are *correlated* correctly. The presence of a positive Lyapunov exponent has the effect of driving the models off the "correct" trajectory in the presence of the slightest of biases, like computer round-off error. For these reasons free-run prediction could not be done for multivariate embedded time series.

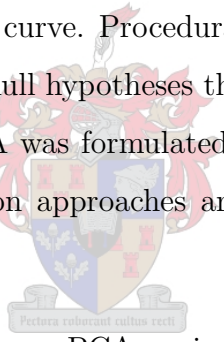
Multivariate uniform embedding is generally superior to univariate embedding. However, the functional relationship of the process variables plays a more decisive role especially for uniform embeddings, as the case of (x, z)-embedding illustrated above. Criteria that determine whether or not two simultaneously channels yield a better reconstruction than either of them used separately are required. The study by Schreiber (2000) lays the theoretical framework for research in this direction.

Probably more important than relationships between variables is the capture of relevant timescales in the embedded data. Non-uniform embedding strategies, though complicated and somewhat lengthy, can achieve this with

remarkable improvement over uniform embeddings using either univariate or multivariate time series. Given that the transition probabilities fluctuate indeterminately in a chaotic attractor, variable embedding strategies offer some hope in resolving this. However, the risk associated with this approach is the loss of the global nature of the derived models.

(c) ***Comparison of dimension reduction methods***

Three dimension reduction methods were compared, namely, principal component analysis (PCA) with and without sphering, and independent component analysis. The probability distribution plots of the (x_j, y_j, z_j) variables are shown in Figure 4.22. Clearly, all of the variables had a different distribution from the Gaussian bell-shaped curve. Procedurally, use of ICA is guaranteed to perform successfully. The null hypotheses that PCA performs just as well as PCA with whitening or ICA was formulated. Some results of using these different latent space projection approaches are shown in Table 4.12. The following were deduced;



(i) For the cases shown, the pure PCA projection method yielded the best performance in the case of MLP predictors. Marginal effect was observed for LSSVM models.

(ii) Significant differences in the correlation/regression coefficients occur mostly for MLP models when ICA is used. A single significant difference was observed for LSSVM models, with $|z| = 2.60 > z_{\alpha/2}$, for a level of significance $\alpha = 0.01$.

The performance of MLP degraded as the complexity of the dimensionality reduction method increased. This resulted in correlation coefficients between the one-step ahead model outputs and actual outputs lower than for PCA–reduced input data. The LSSVM were generally somewhat robust to different pre-processing approaches, both in terms of the resulting performance and the correlation coefficients of the one-step ahead predictor outputs and actual outputs.

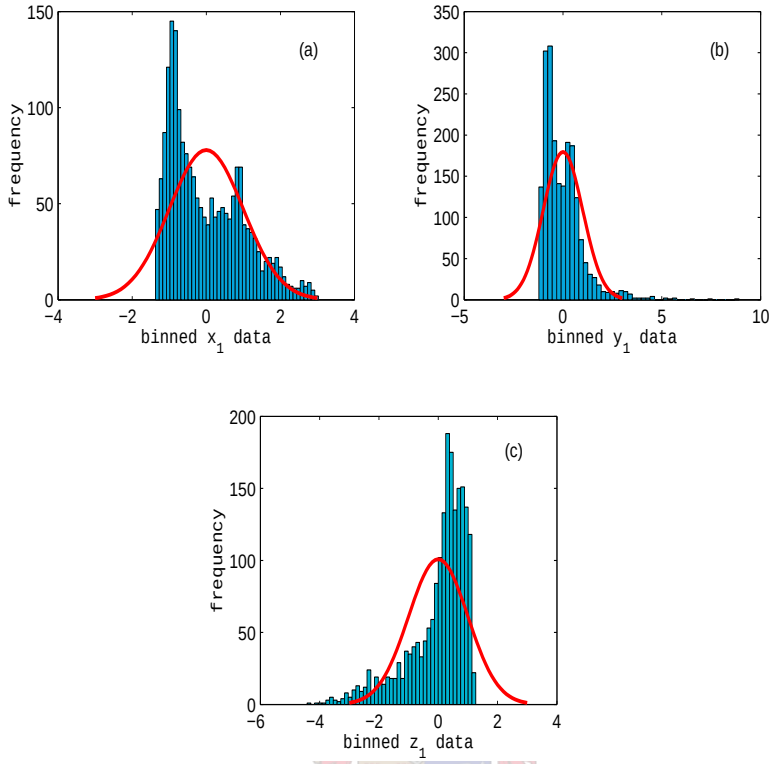


Figure 4.22: Histogram plots of the x, y, z data from the coupled CSTR after normalization, each plot showing a superimposed normal fit on the data

(c) ***MLP or LSSVM – which performs better?***

To determine the relative merits of using MLPs over LSSVMs or vice versa, direct comparison were made of the mean square error and the significance in differences of the correlation coefficients for the different models. Figures 4.23 and 4.24 are plots of the performance and significance tests for the uniform and non-uniform embedding embedding strategies.

The following analysis follows from results in Figures 4.23 and 4.24;

- In the case of uniform embedding, MLP models out-performed LSSVM models in all but one instance, (x, y)-model, with respect to the mean square error criterion.

- The estimated correlation coefficients were significantly different in all cases considered. No discernible pattern observed as two of the cases yielded a better correlation coefficient for the LSSVM whilst in the rest the MLP correlation coefficient was better (wrt. uniform embedding).
- For non-uniform embedding, MLP model performed generally better than LSSVMs. However, when three variables were used in the reconstruction both classes exhibited similar performance.
- Estimated correlation coefficients for LSSVMs were significantly better than corresponding values for MLPs, with the exception of the (x, y)-model that has showed an opposite trend (wrt. non-uniform embedding).

From a consideration of the foregoing it is difficult to conclude which model class performs or generalizes better than the other with respect to modeling of data from a coupled CSTR system. The effect of the prediction time step in the resulting models (alternatively, the effect of autocorrelation in the variables) was investigated and typical results obtained are presented in Figure 4.25. Similar trends were observed for the other embedding strategies.

In general, it was observe that as the prediction time step p was increased (alternatively, as autocorrelation between variables decreased), the LSSVM performed better than the MLP. It is not clear why this is so but MLPs are known to make use of autocorrelations training data. As the correlations decrease the performance of MLPs degraded faster than for LSSVMs.

On the basis of results obtained it can be concluded that in the modeling of a coupled CSTR system MLPS performed better than LSSVMs. However, LSSVMs were less sensitive to preprocessing of data than MLPs. It should remarked though that LSSVMs have been used with success where MLPs have failed, for example, the two spiral classification problem (Suykens, 2001). Re-

search into LSSVMs is still ongoing and with further theoretical developments LSSVMs may eventually approach or surpass MLPs in terms of performance in certain applications. The current major drawback of using support vector machines is the computational time it takes to optimize the regularization and kernel parameters.

4.8 Comparison of $d_c(\varepsilon)$ Estimates for Different Embedding Strategies

The correlation dimension is an estimate of the interpoint probability distribution function of a set in state space and gives an estimate of the effective degrees of freedom of a dynamical system. An embedding dimension must always be greater than the fractal dimension to avoid the attractor "filling" the embedding space. Figure 4.26 indicates that the embedding strategies employed result in more similar estimates for the correlation dimension, $d_c(\epsilon_0) \approx 2.5$. Using the original state space, however, indicates that the trues attractor has a dimension in the range $2.5 - 3$. Also, the true correlation dimension estimate is defined at lower scales than most of embedding strategies employed, with the only exception of non-uniform multivariate. It is apparent that non-uniform multivariate embedding captures better the intrinsic dynamics at much lower scales that univariate uniform embedding. Therefore, in relation to the modeling results presented earlier in section 4.7.2, it can be postulated that the improved performance of non-uniform embedding is due to the embedding's ability to capture small scale dynamics than other strategies. This further confirms results obtained in the local linear approximation models in Figure 4.20.

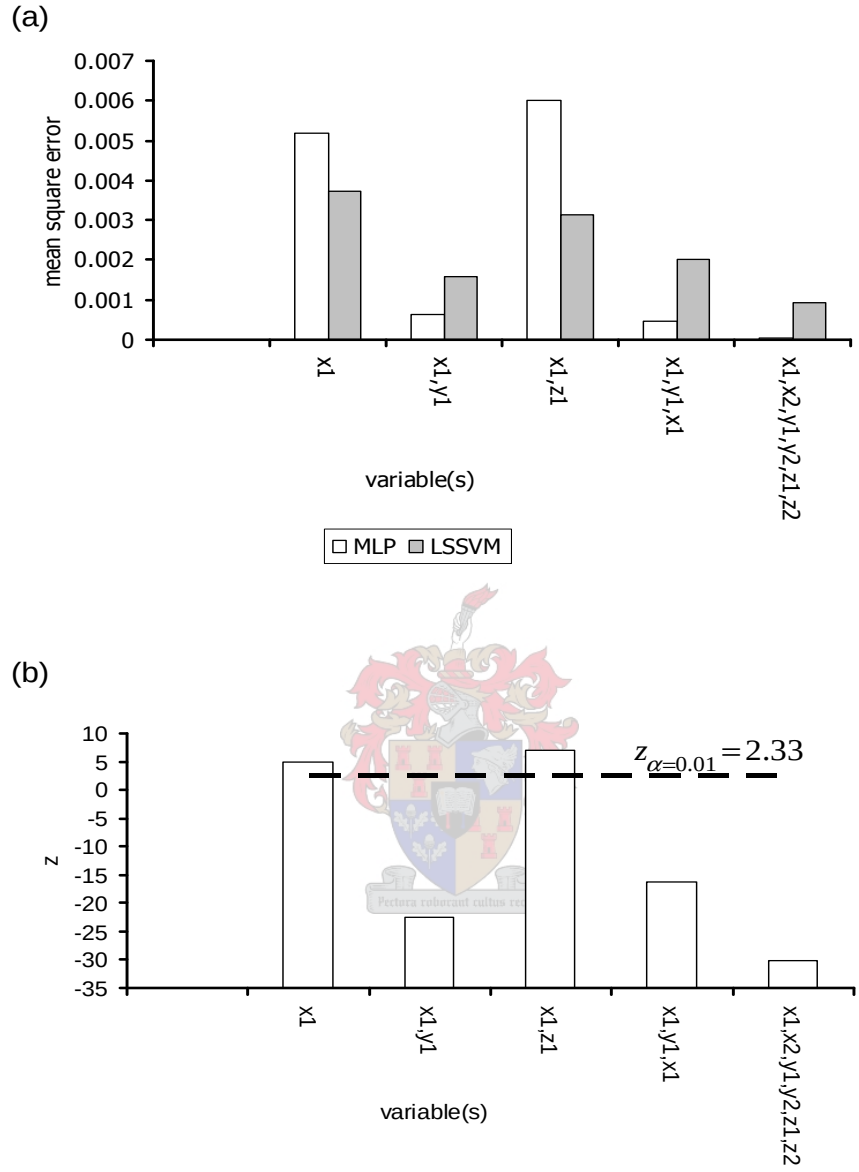


Figure 4.23: Comparison of the MLP and LSSVM models after reconstruction of the state space using *uniform* embedding. (a) Plot of MSE for the respective models and variables. (b) Test of significance in the difference of $\hat{\rho}$ computed from MLP and LSSVM model outputs. The dashed line is the critical value of the z statistic above which $\hat{\rho}$ for an LSSVM model is significantly better than the corresponding MLP model.

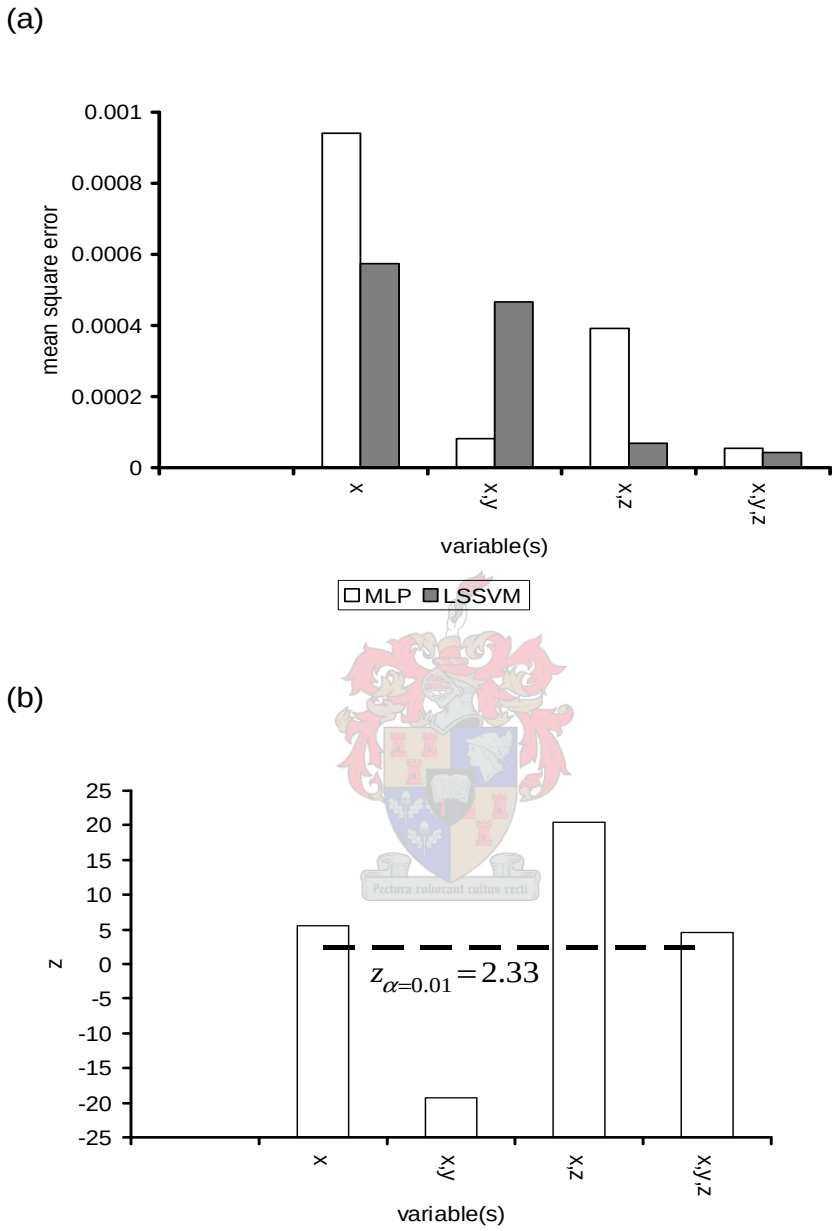


Figure 4.24: Comparison of the MLP and LSSVM models after reconstructing the state space using *non-uniform* embedding. (a) Plot of MSE the respective models and variables. (b) Test of significance in the difference of $\hat{\rho}$ computed from MLP and LSSVM model outputs. The dashed line is the critical value of the z statistic above which $\hat{\rho}$ for an LSSVM model is significantly better than the corresponding MLP model.

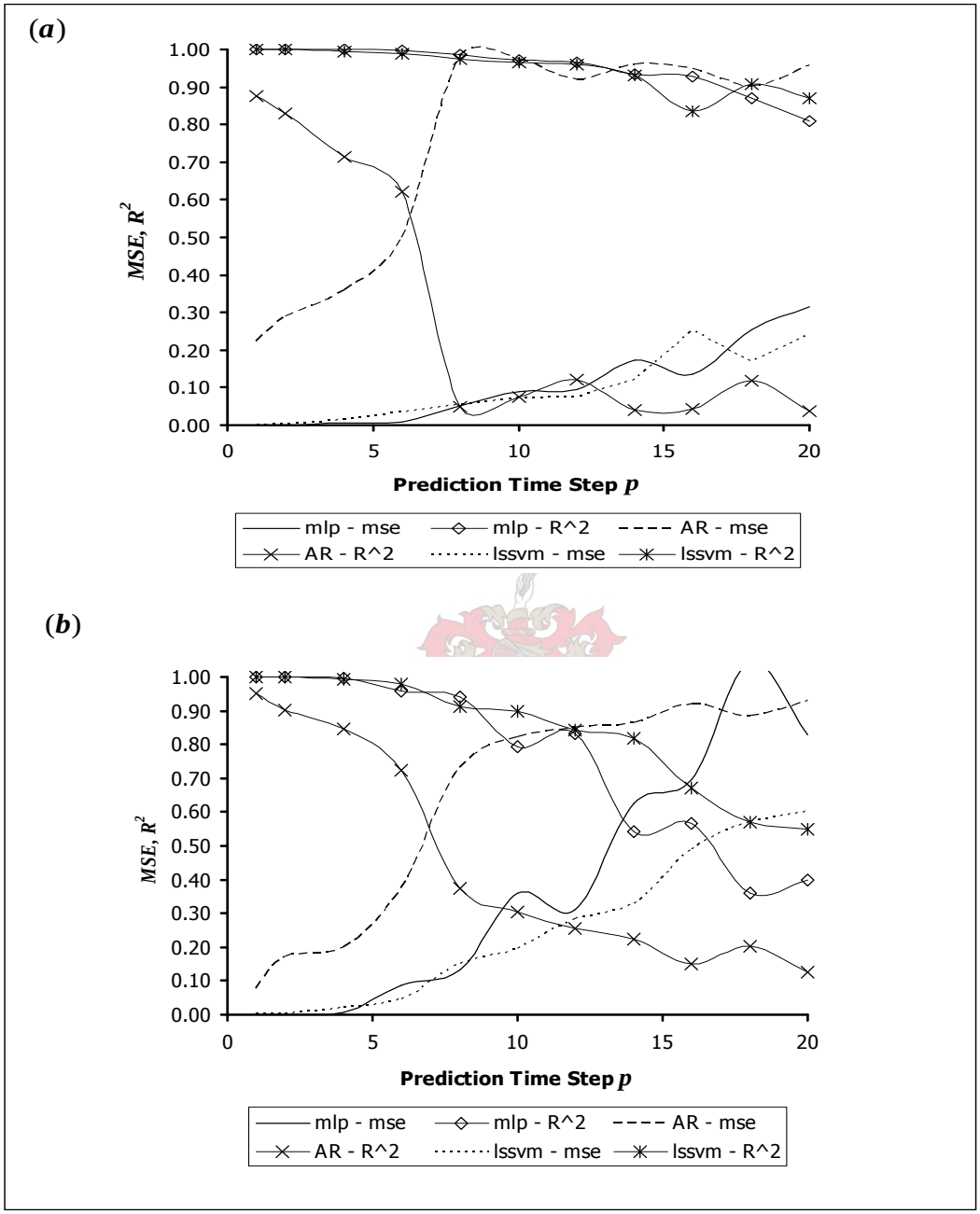


Figure 4.25: Performance of different model classes with variation in prediction time step. (a) $x(t+p) = f(\mathbf{x}_t)$, $\mathbb{R} \leftarrow \mathbb{R}^6$, where $\mathbf{x}$ is non-embedded vector of the underlying 6 variables. (b) $x(t+p) = f(\mathbf{x}_t)$, $\mathbb{R} \leftarrow \mathbb{R}^{35}$, where $\mathbf{x}_t$ is the embedded vector using the variables, $x_1, y_1,$ and z_1. AR refers to an equivalent linear autoregressive model shown for comparison.

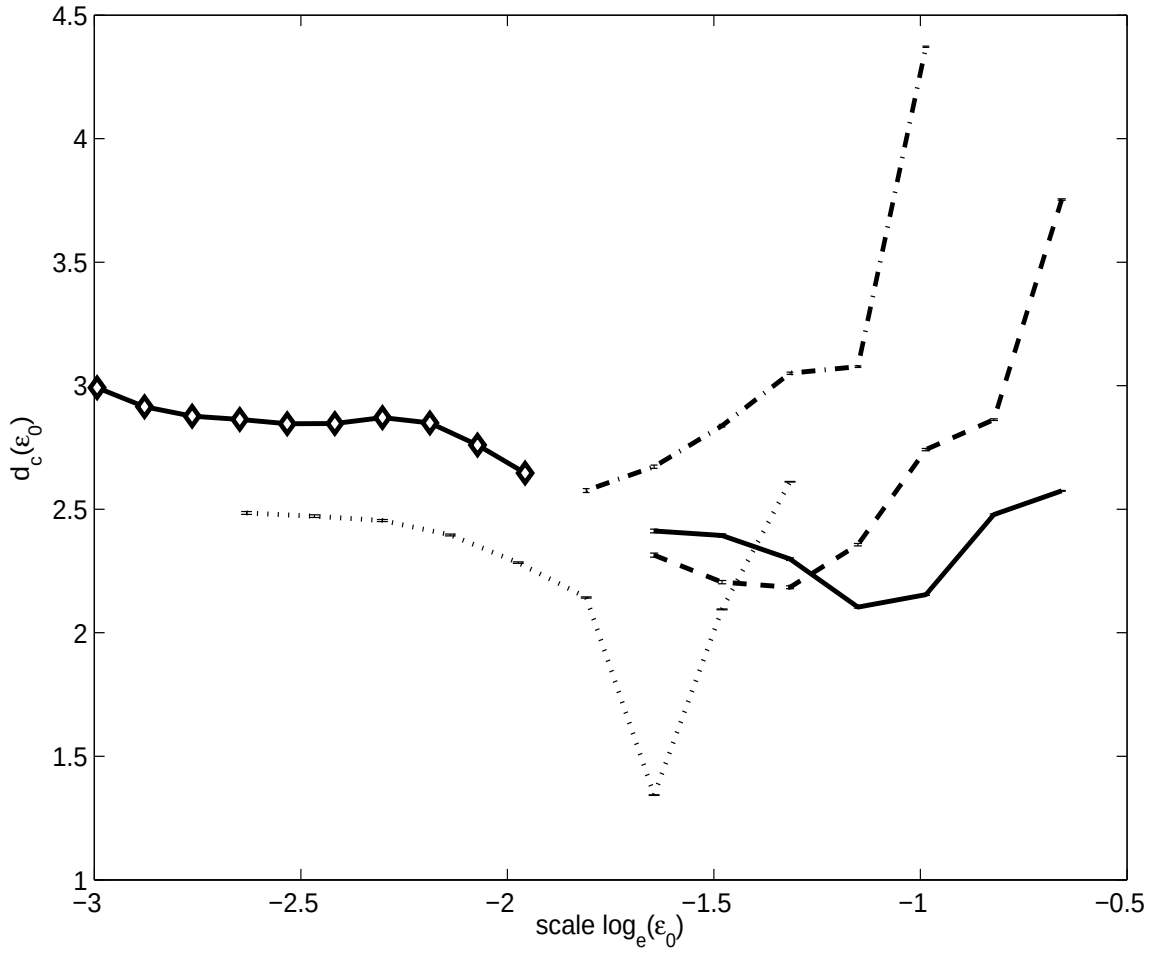


Figure 4.26: Correlation dimension estimates for different embedding strategies.

KEY

solid curve – Takens' uniform embedding of x_j;

dashed – non-uniform embedding of x;

dash-dot – uniform trivariate (x, y, z)-embedding;

dotted – non-uniform trivariate (x, y, z)-embedding;

solid +diamonds – trivial embedding of underlying variables $\in \mathbb{R}^6$.

4.9 Concluding Remarks

With respect to the detection of nonlinearity in time series signals, surrogate data that preserved cross-correlations among observed signals improved the robustness of the test. Ignoring cross-correlations risks spurious rejection of the null hypothesis of an underlying linear process, possibly observed through a nonlinear measurement function. Use of the correlation dimension as a discriminating statistic, though simpler and easily interpreted, must be accompanied by other statistics for unequivocal characterization of time series data. These other statistics may include entropy and Lyapunov exponents. It should be noted, however, that algorithmic implementations of these other statistics are inherently less reliable and cannot be used separately.

In general, use of multivariate data from a coupled CSTR in the reconstruction improved the model performance compared to scalar embedding. However, the choice of channels used in the reconstruction and the embedding strategy was critical. In particular, use of certain variables did not improve significantly the performance of the model compared to the univariate reconstruction. This was particularly evident in multilayer perceptrons. Also, use of non-uniform embedding strategies improved model predictability than uniform embedding. This was irrespective of the choice of process variables used in the embedding.

Analysis of correlation dimension estimates for all the embedding strategies gave a similar estimate. However, in the case of non-uniform embedding it was observed that the estimate was defined at much smaller scales. Hence, it was concluded that use of multivariate non-uniform embedding enabled the dynamics defined over high frequencies to be captured in the reconstruction and, consequently, better predictive model.

Chapter 5

Case Study: System Identification of Industrial Flotation Plants

5.1 Process Description

To demonstrate the usefulness of multivariate time series analysis and also use of the least-squares support vector learning in an industrial environment, modelling of the dynamic behaviour of a real data set from a lead (Pb) and zinc (Zn) metal flotation circuit was investigated. The data was taken from a flotation plant which treats Pb-Zn–bearing ore slurry ground to a required size distribution in tumbling mills. Pb and Zn concentrates are separately produced in the flotation circuit using both conventional and proprietary technology. An important concern is control of losses of lead and zinc in the tailings to the least minimal possible levels or, alternatively, maximize recovery of the two metals in the concentrates. In any flotation process, recovery is affected by many process variables that include pulp density, aeration rate, and grind size amongst other factors. It was assumed that variation of the observed values of iron in the tailings, $Fe_T(t)$ could be defined by

$$Fe_T(t+1) \quad = \quad f\big(\mathbf{x}_{Fe}(t), \mathbf{x}_{Pb}(t)\big) \tag{5.1}$$

where $f, \mathbf{x}_{Fe}(t), \mathbf{x}_{Pb}(t)$ are the model function, the reconstructed vectors from observed time series of Fe_T and Pb_T respectively.

Figure 5.1 shows the time series plots of Pb and Zn in the tails stream observed over a period of 632 time steps.

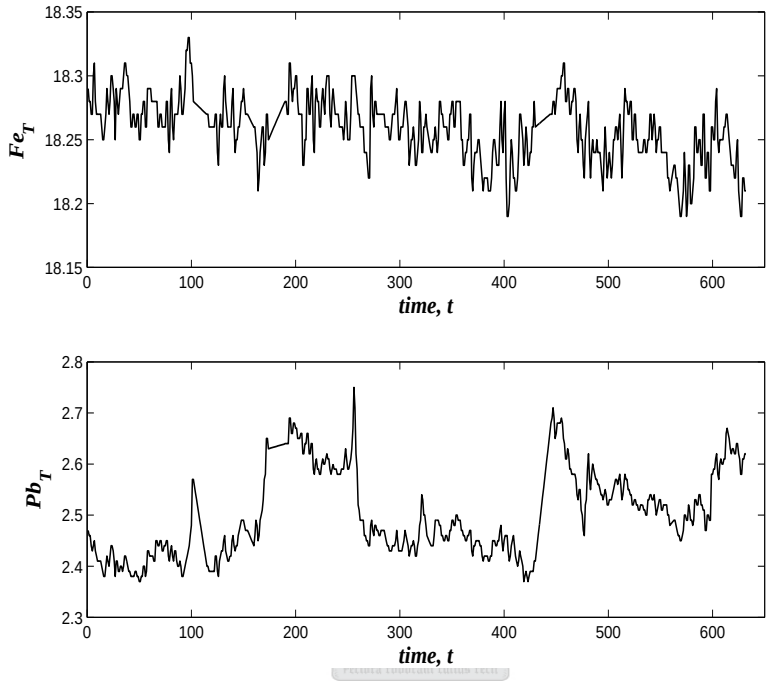


Figure 5.1: Variations of Fe_T and Pb_T with time in the tailings from a Pb-Zn flotation process as observed

5.2 Data Preprocessing

Recalling an earlier discussion, the data to be analyzed must justify the use of nonlinear time series analysis techniques is before such tools are applied. Detection of nonlinear behaviour (or, more correctly, inadequacy of a purely linear description) is commonly done using the method of surrogate data. Linear trends were observed in the time series plots of the observed data, particularly in the plot of iron val-

ues, Figure 5.1. The data was detrended to remove these linear structures. The detrended data plotted in Figure 5.2 were used for detection of possible non-linear structures using surrogate data analysis and estimating the nonlinear model f in equation 5.1

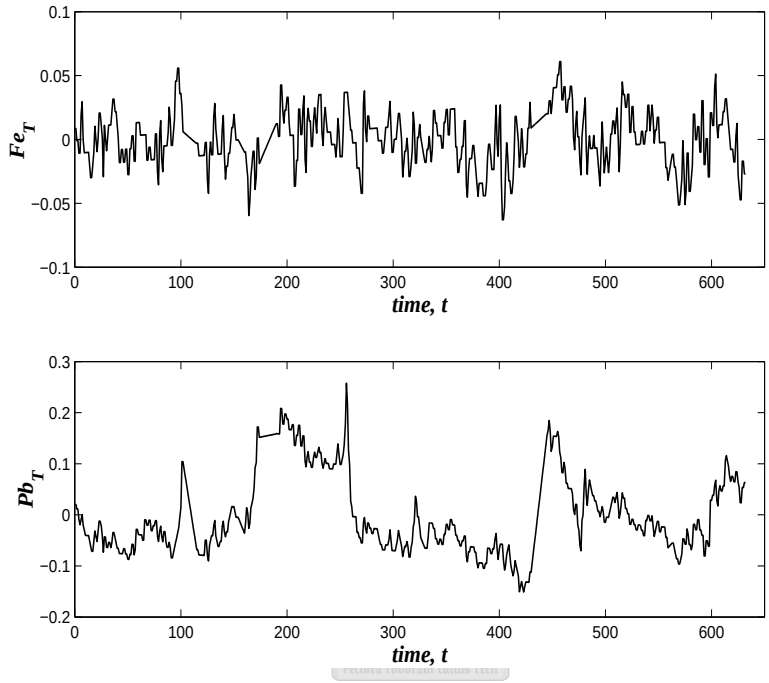


Figure 5.2: Time series plots of the variation of Fe_T and Pb_T in the tailings stream after detrending.

5.3 Results of Surrogate Analysis

Two sets of surrogate data were generated for the each of the iron and lead values; (i) using a single times of the detrended values, and (ii) using both detrended time series. As alluded to previously, the use of multivariate data in surrogate data analysis potentially improves the identification of nonlinearity by preserving cross-correlations existing between various channels in the surrogate generation algorithm.

Judd's implementation for correlation dimension estimate as a function of viewing scale ($d_c(\varepsilon_0)$) was used as the discriminating statistic.

The results of the surrogate analysis tests are shown in Figures 5.3 and 5.4. Embedding parameters of $d_e = 10$ and $T = 1$ were used for both the original data and surrogates. Higher embedding dimensions could not be used because of the small time series lengths. It must be noted, however, that higher d_e values are only necessary in instances where preliminary results show "filling" of the space for selected embedding dimensions, which was not the case in this instance.

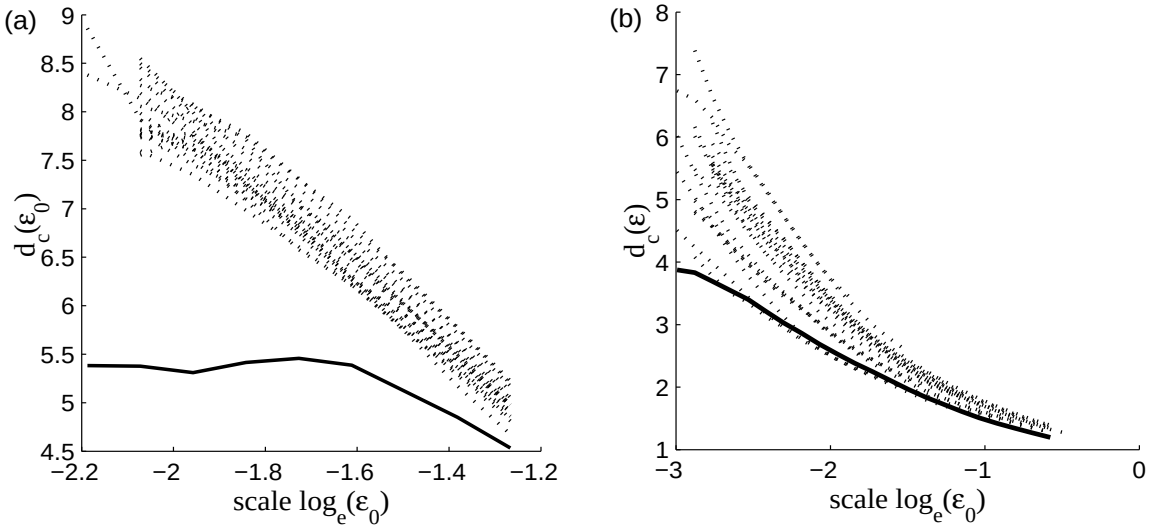


Figure 5.3: (a)Testing for nonlinearity using Fe_T values in the surrogates generation. A separation between the estimates of d_c^{data} and d_c^{surr} particularly at smaller scales is evident. (b) Testing for nonlinearity using Pb_T values in the surrogates generation. The d_c estimates for the data overlap with those of the surrogates and discounting presence of an underlying nonlinear process.

The plots from the surrogate analysis using Fe_T, Figures 5.3(a) and 5.4(a), show a clear separation of the $d_c(\varepsilon)$ estimates before and after preserving cross-correlations in the surrogates. Moreover, a persistently constant scaling region at smaller scales for the correlation estimate of the data suggests a fractal object

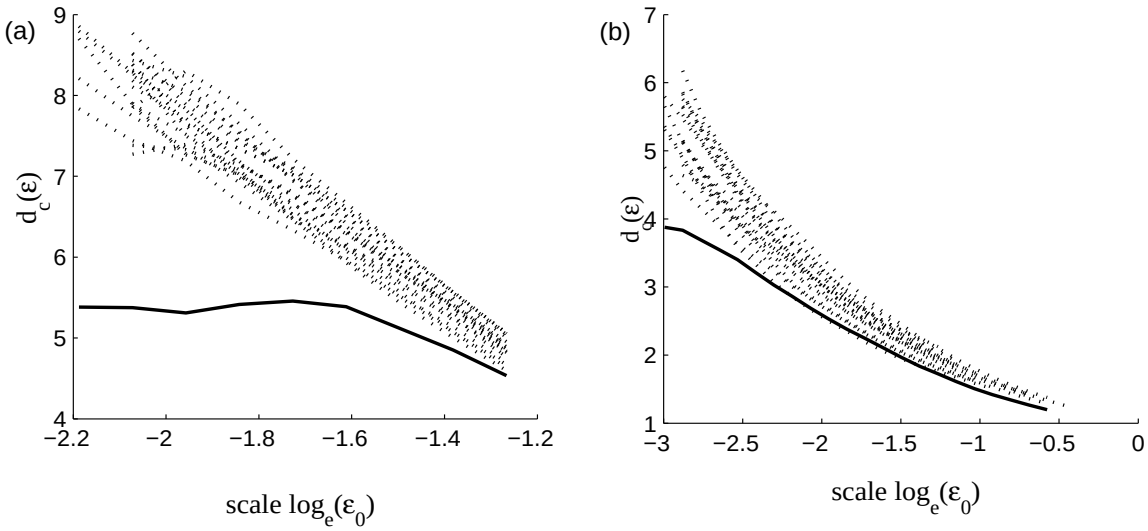


Figure 5.4: Multivariate nonlinearity testing for the flotation data. Here the surrogates were generated using Fe_T with cross-correlations from Pb_T having been taken into account. Separation is still evident at the smaller scales whilst an overlap exists in the larger scales (b) The d_c estimates for the Pb_Tdata still overlap with those of the surrogates. Notice, however, that the trend in the smaller scales suggest a separation occuring.

with a dimension of about 5.4. In light of the inherent bias in Judd's algorithm for d_c values greater than 4 (Judd, 1992) a cautious approach was required in the interpretation of the results of the correlation dimension. The data suggested low-dimensional determinism and, therefore could be exploited to fit nonlinear models. Specifically, the dimension suggested the flotation process evolved in a phase space with effective number of degrees of freedom < 6. However, the actual value was difficult to ascertain from the limited time series data and algorithmic bias in the d_c estimate obtained.

Results of surrogate tests based on Pb values are shown in Figures 5.3(b) and 5.4(b). Clearly, the null hypothesis of a static monotic nonlinear transformation

of linearly filtered noise could not be rejected. However, it was observed that there were indications of a separation between the surrogate data and Pb_T values in the smaller scales of the plots. This was particularly so when cross-correlations between the surrogates of Pb_T and Fe_T were preserved, Figure 5.4(b). This was contrary to what would have been expected. As pointed out in the analysis of the coupled CSTR system, the surrogate generating algorithm may not be constraining both surrogates correctly. Also, the "pivotalness" of the correlation dimension estimate when cross-correlations are preserved is in doubt.

It was concluded that there were structures in the data that could not be attributed to purely random noise, although the evidence was marginal for Pb_T values. More information especially on the state of process operating conditions may have clarified the results.

5.4 Fitting Nonlinear Models

Motivated by the results from the surrogate analysis, nonlinear models were fitted to estimate f in equation (5.1). Bearing in mind that any embedding dimension d_e at least greater than the fractal dimension can be used for state space reconstruction and, that the limited data length risks unreliable estimates of the embedding space or time delay, the reconstruction was resolved as follows: the time delay was fixed at $T = 1$ sampling time steps. For each of the embedding window values in the set $[10; 20; 30; 40; 50]$ models were constructed using both LSSVMs and MLPs. One-step ahead predictions were performed using a model trained on an embedded vector space of 500 points and validated on the rest of the points not used in the training. In the case of iterative or free-run predictions, two approaches were considered. "Honest" iterative predictions were defined on models parameterized using training set of 500 points and validated on the remainder of the data not used in

training the model. "Dishonest" iterative predictions were defined on models parameterized using the entire data set and then predicted iteratively over a specified prediction horizon T_w using the first embedded vector in the reconstructed trajectory. The reason for such a procedure and distinction was to test the robustness of the models, both in their construction and in predicting possibly chaotic phenomena. In general, it is observed that poorly constructed nonlinear models may yield good one-step ahead performance but collapse within a short time horizon under an iterative prediction scheme. In particular, under iterative prediction using the *training* data they tend to fixed points, implying that virtually the fitted model did not capture the underlying dynamical behaviour. Long-term prediction is not possible for chaotic phenomena because of the existence of a positive maximal Lyapunov exponent. A measure of how good a model performs in the long-term is, therefore, its generalization ability. Such a parameterized model traces future trends whilst remaining within a variance similar to that exhibited by the system data. An "honest" iterative prediction helps to assess the model's long-term generalization ability. Tables 5.2-5.1 and corresponding figures are summary results obtained in the modelling.

Table 5.1: One-step ahead performance statistics for a predictor built from the bivariate time series from the flotation plant. For reasons explained in the main text, negative R^2 indicates zero correlation.

Embedding	LSSVM Modelling				MLP Modelling			
window	MSE	R_c^2	Z	z	MSE	R_c^2	Z	z
10(5)	1.89E-04	0.99	-0.28	0.74	7.39E-04	0.1208	0.5995	0.22
20(10)	2.48E-04	0.9481	-0.25	0.6459	3.10E-03	-2.7587	0.1328	-0.27
30(15)	2.64E-04	0.9764	1.67	0.5632	1.40E-03	-0.7915	0.4302	0.57
40(20)	2.80E-04	0.8349	0.67	0.5123	1.20E-03	-0.371	0.1660	-2.49
50(25)	3.05E-04	0.7686	1.51	0.4135	9.79E-04	-0.338	0.5303	0.425

Table 5.2: Summary of LSSVM modelling results of the flotation process. Iterative prediction statistics are for the "dishonest" case.

Embedding	One-Step Predictor		Iterative Prediction($T_w = 100$)			Iterative Prediction($T_w = 300$)		Iterative Prediction($T_w = 500$)	
window	MSE	R^2	Z	MSE	R^2	MSE	R^2	MSE	R^2
10	1.82E-04	0.6575	1.1364	3.35E-04	–	–	–	–	–
20	2.27E-04	0.5475	0.981	9.62E-04	–	–	–	–	–
30	3.18E-04	0.3634	0.7472	1.46E-07	1.00	1.01E-06	0.9969	0.2103	3.41E-04
40	3.47E-04	0.3546	0.7466	1.96E-08	1.00	2.82E-08	0.9999	1.04E-07	0.9998
50	4.81E-04	0.1077	0.5432	4.62E-09	1.00	9.63E-09	1.00	1.21E-08	1

Table 5.3: Summary of MLP modelling results of the flotation process. Iterative prediction statistics are for the "dishonest" case. Note the shorter "dishonest" iterative prediction horizon compared to the LSSVM modelling results above. Negative R^2 values are an artefact of outliers in the data and are generally assigned a zero values.

Embedding	One-Step Predictor		Iterative Prediction($T_w = 25$)			Iterative Prediction($T_w = 40$)		Iterative Prediction($T_w = 60$)	
window	MSE	R^2	Z	MSE	R^2	MSE	R^2	MSE	R^2
10	1.10E-03	-1.0627	0.5710	3.08E-04*	0.072*	–	–	–	–
20	2.50E-03	-4.0271	0.1682	2.82E-04	-0.1769	–	–	–	–
30	9.52E-04	-0.9056	0.3517	9.28E-06	0.9672	9.12E-05	0.5938	–	–
40	1.00E-03	0.2638	0.5214	7.50E-06	0.9503	5.68E-04	-2.7178	–	–
50	7.29E-04	-0.2612	0.4669	3.46E-06	0.9745	6.90E-06	0.9622	3.04E-05	0.9147

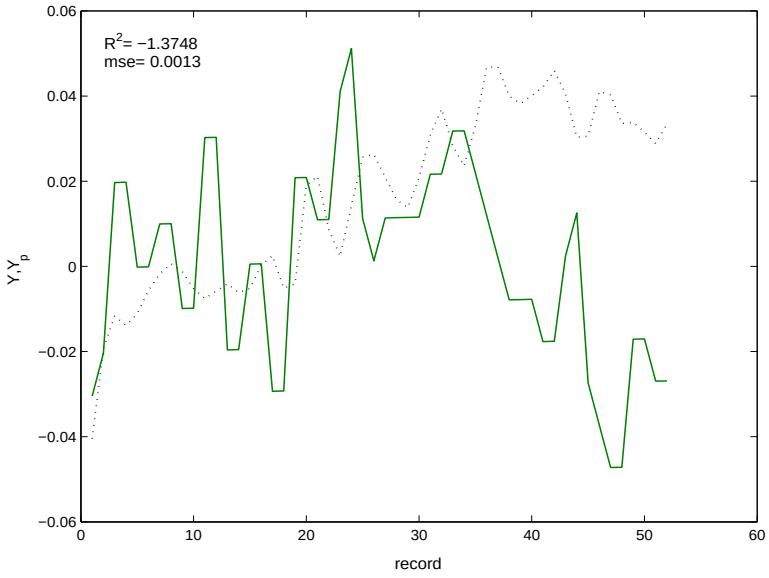


Figure 5.5: LSSVM iterative "honest" prediction with $d_e = 30$. The negative R^2 value is due to outliers, observed mostly after about 30 time steps.

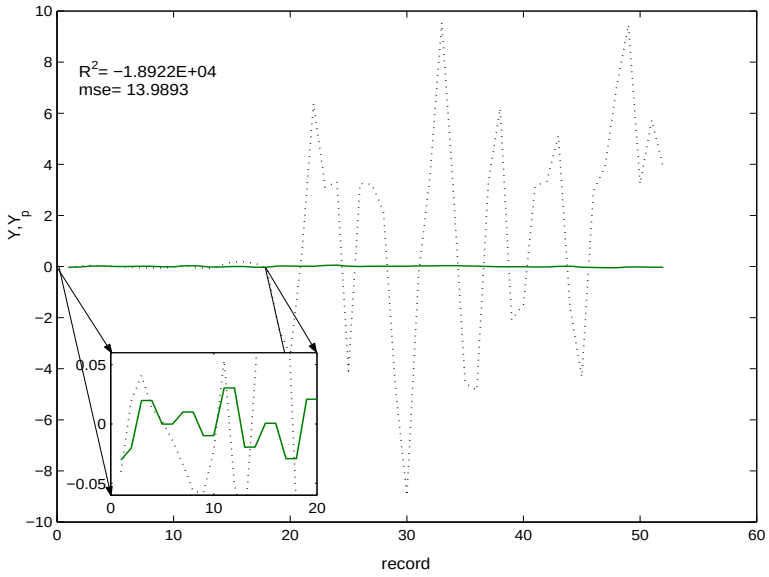


Figure 5.6: MLP iterative "honest" prediction with $d_e = 30$. The nonsensical R^2 is due to the wild fluctuations observed after about 10 time steps. Strictly speaking, the R^2 value must be calculated only in that region.

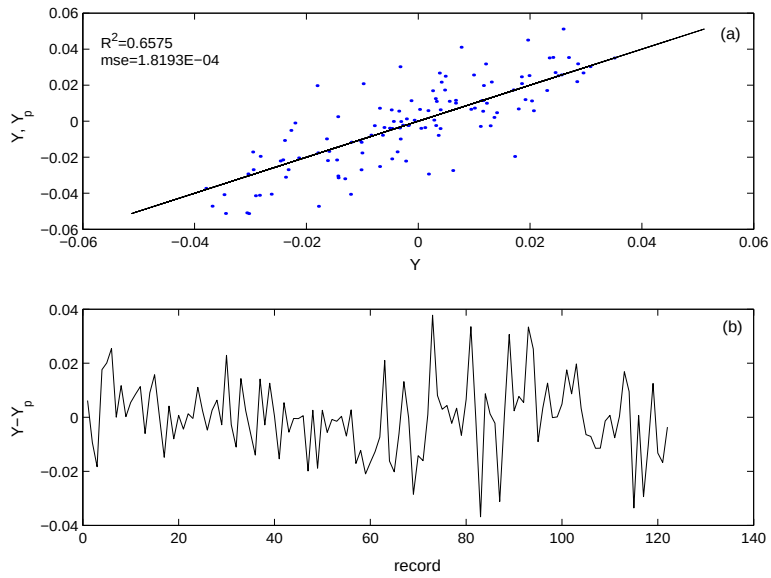


Figure 5.7: LSSVM one-step ahead predictor results with $d_e = 10$ (a) regression statistics between the actual data (-) and predicted values ($\cdot$) (b) residual plot of the difference between actual (Y) and predicted (Y_p) values.

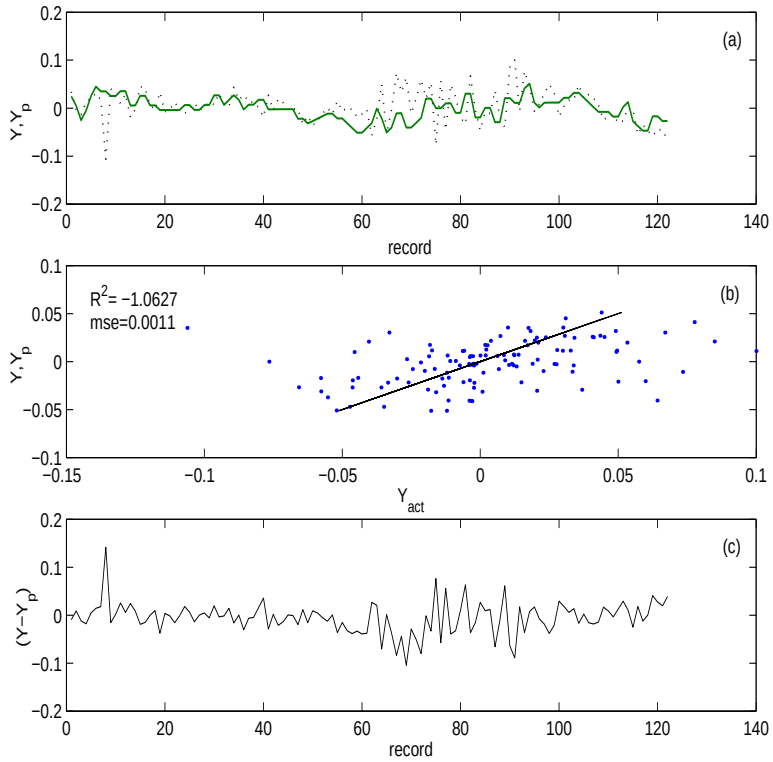


Figure 5.8: MLP one-step ahead predictor results with $d_e = 10$ (a) time series plots of the predicted(:) and actual values(-) (b) regression statistics between the actual data (-) and predicted values (·) (c) residual plot of the difference between actual (Y) and predicted (Y_p) values

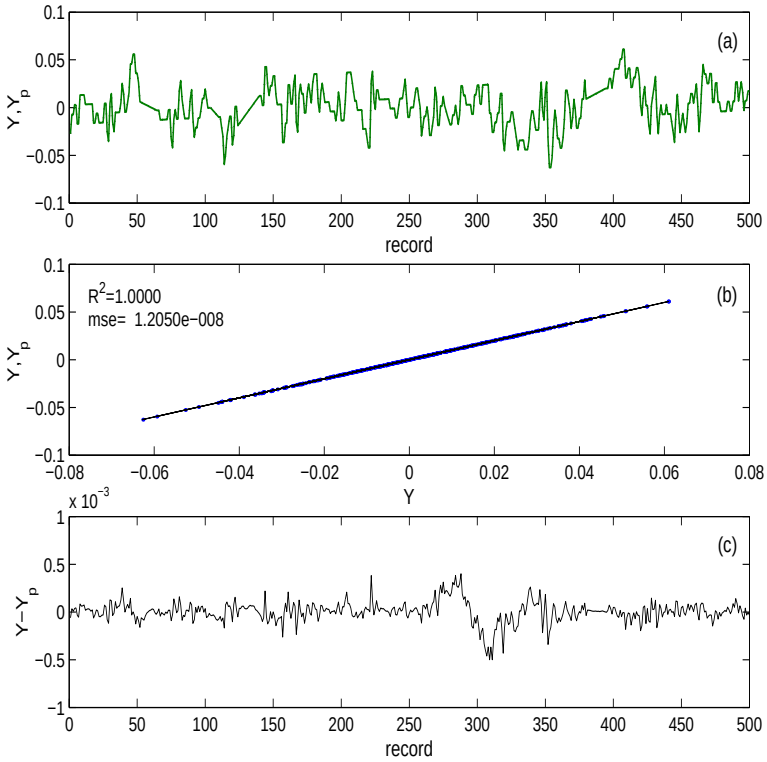


Figure 5.9: LSSVM "dishonest" predictions for a prediction horizon $T_w = 500$ (a) time series plots of the predicted(:) and actual values(-) (b) regression statistics between the actual data (-) and predicted values (·) (c) residual plot of the difference between actual (Y) and predicted (Y_p) values

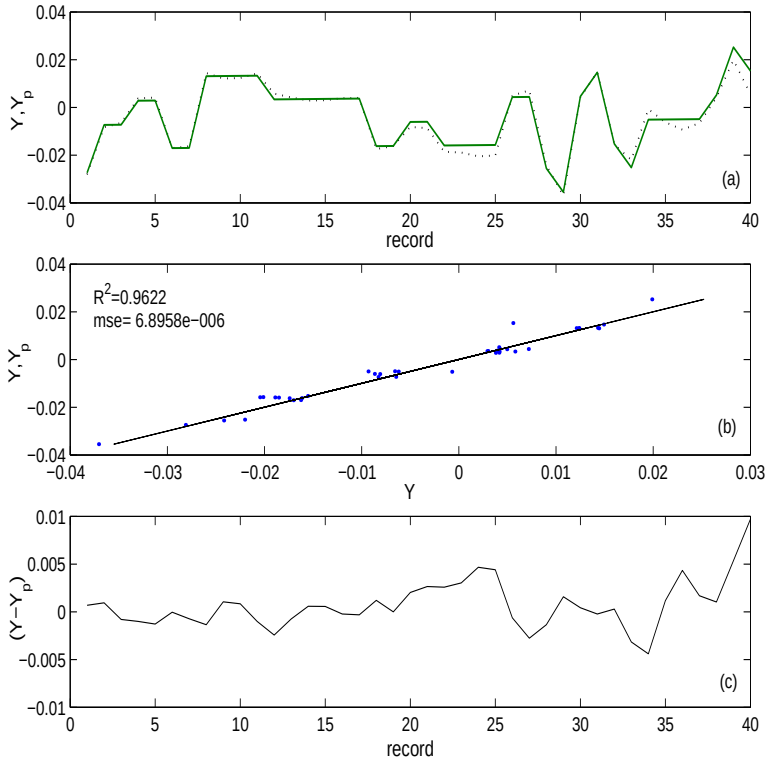


Figure 5.10: MLP "dishonest" predictions for a prediction horizon $T_w = 500$ (a) time series plots of the predicted(:) and actual values(-) (b) regression statistics between the actual data (-) and predicted values ($\cdot$) (c) residual plot of the difference between actual (Y) and predicted (Y_p) values

5.5 Discussion and Concluding Remarks

The LSSVM models fared better than MLP models in all the cases shown. For the one-step ahead models, positive regression coefficients were obtained for the LSSVMs. However, the R^2 values decreased with increasing embedding dimensions. This is possibly due to sparseness of the reconstructed space in higher embedding dimensions. Almost all MLP models except one showed had negative R^2 values calculated over the entire length of the validation data. In other words, none of the fitted MLP models was able to learn the time evolution of the data. Negative

correlation coefficients are a result of many outliers in the regression plot indicating that using average values of past observations to predict the next value is better than using of the fitted model. It is usual to assign these values a zero R^2 value. The regression coefficients are shown as obtained to give an indication of how worse the fitted model was than a corresponding model that uses an average of the past values.

In "dishonest" iterative predictions, both model classes' performance improved with increasing d_e. This is expected because for increasing d_e the models obtained approach a local linear approximator. The effect was particularly exceptional for the LSSVMs, which gave perfect predictions ($R^2 = 1, MSE \ll 0$) in some cases. More significant is the fact that iterative predictions of LSSVM models were robust in retaining the structure of the trained data over a much broader prediction horizon than the MLPs. This is in agreement with the structure and construction of the support vector machine. The predicted value is a weighted sum of its distance from support vector nodes. In the case of least squares machines, each of the training points is a support vector and therefore affects positively future points that lie within its "basin of attraction". This is in spite of the sensitivity to initial conditions of chaotic phenomena. Sparse approximation is possible where only the effective support vectors are retained. However, this has dramatic effects on the performance of the fitted models.

The "honest" iterative predictions further confirm the better generalization ability of LSSVMs. Figures 5.5 and 5.6 clearly show the LSSVM's ability to generalize over a wider prediction horizon than MLPs. The MLP's predicted trajectory path escapes the region of space explored by the data after 20 time steps whilst for the LSSVM the predicted trajectory remains restricted within the limits of the underlying system.

Including additional information from another observed variable improved the performance of the LSSVM one-step ahead predictors as indicated in Table 5.1.

Iterative predictions could not be performed on the bivariate data for reasons mentioned earlier (i.e., correlation problem introduced by separately optimized models for each variable). It is asserted that the similar conclusions hold as in the univariate case,i.e., LSSVMs have better generalization capabilities than MLPs.

Bivariate embedding did not give significant improvements of the models' performance. One-sided statistical tests for the differences in correlations of the predicted and actual values were performed under the null hypothesis – $H_0 : \hat{\rho_0} = \hat{\rho_i}$ against the alternative hypothesis $H_1 : \hat{\rho_0} < \hat{\rho_i}$, with $\hat{\rho_0}$ being the estimate of the correlation coefficient for a model built using univariate time series and $\hat{\rho_i}$ the corresponding equivalent multivariate-based model (equivalent in the sense of equal embedding dimension). The null hypothesis could not be rejected at the $\alpha = 0.01$ significance level. However, at higher embeddings dimensions $d_e \geq 20$ the least-squares support vector machine showed a single instance ($d_e = 30$) where H_0 was rejected. The results are as expected given that the surrogate tests hardly detected evidence of nonlinearity in Pb_T data, Figure 5.3(b) and 5.4(b).

In conclusion, therefore, in the modelling of plant data of limited size from a flotation process, LSSVM models performed better than MLPs, both in terms of least forecast errors and generalization ability. This was in sharp contrast to results obtained in the modelling of a coupled CSTR system. Probable reasons for this are the much limited training data lengths from the flotation process, poor autocorrelations in the data, and larger variances in the plant data compared to the simulated data. Further work is needed to substantiate this. The modelling of the coupled CSTR system and other known dynamical map could be done using different data sizes, sampling intervals, and adding various levels of noise. It is expected that such an investigation would give conditions or assumptions under which one learning algorithm performs better than the other.

Chapter 6

Spatiotemporal Analysis

"You can only find truth with logic if you have already found truth without it"

GK Chesterton

Reaction and conversion processes in chemical reactive systems are effected and affected by transport processes occuring at both large and microscopic levels. The transport process are induced by concentration gradients existing in the systems. Typical examples of such systems include reaction-diffusion processes in hydrometallurgical processes, adsorption in surface chemistry processes, metal dissolution and deposition in electrochemical engineering, ion-exchange processes, etc. A proper dynamical analysis of these processes is appropriately described by field theory concepts using partial differential equations (PDEs). In practice, however, the analysis of such systems is simplified by assuming homogeneity and taking measurements at a single spatial point. Alternative modelling strategies consider averages over the spatial behaviour (lumped parameter modelling).

To illustrate the broader context which motivates spatiotemporal time series analysis consider the use of carbon in various chemical processing plants, for example, gold adsorption and de-colouring processes that use fluidized carbon beds of specified grain size distribution in solution. Each of the carbon particles is com-

posed of numerous interconnected irregular pores. The mass transfer of species to and from the granular carbon surface can be described by three diffusion mechanisms, depending on the distribution of pore sizes in a given particle. Knowledge of such a distribution is very difficult or impossible to obtain. Representing the carbon particle by a continuous single pseudo-homogenous phase allows one to define an *effective diffusion coefficient* over the pore size distribution. This aids the process engineer in the development of practical models for the understanding of system behaviour.

However, despite the simplification of lumping parameters, the resulting PDE models are of very high order and do not allow for tractable models to be obtained with ease. Even though proper analysis of such models can be done numerically given the computer advances to date, such models are of little use in controller or reactor design. Also, the simplification introduced by lumped parameter models potentially risks under-resolution through failure of the lumped model to capture certain dynamical behaviour ("small causes also give rise to large effects", according to low-dimensional determinism). Therefore, it is logical to attempt dynamical analysis of these high-dimensional systems using concepts and tools from low-dimensional nonlinear dynamics. There is evidence to suggest that such an approach can be successful (Ørstavik *et al.*, 1998). In this chapter, reconstruction of spatially extended systems was investigated using the coupled logistic map lattice as a candidate system. The effect of including spatial as well as temporal information in predictive modelling is investigated. Without loss of generality, the least-squares vector machines with Gaussian radial basis kernels was used in fitting models on observed data from the system.

6.1 Reconstruction and Prediction of a CML

6.1.1 Spatially extended systems

A spatiotemporal or spatially extended system is a collection of subsystems in a given spatial configuration, which extends infinitely far in all directions to give an infinite number of interacting state variables coupled into one large system (Cross and Hohenburg, 1993; Diks *et al.*, 1997). Figure 6.2(a) is a schematic illustration of the temporal evolution of a one-dimensional lattice system. The time evolution at each site or lattice is according to some dynamical rule that, for simplicity, is usually assumed to be homogenous across the entire lattice system.

The Coupled Map Lattice

The coupled map lattice (CML) was introduced by Kaneko (1989a,b) and is a dynamical system with discrete time, discrete space, and continuous state. In contrast, a partial differential equation has continuous state, continuous time and continuous space. A characteristic feature of coupled map lattice is that the underlying physical space is a discrete structure or lattice Ω. Points within the lattice are called sites ω and can be finite or infinite. A local state space X_ω with an uncountable number of elements is defined at each site.The state space $\mathcal{M} = \prod_{\omega\in\Omega}$ of a CML is a product of local state spaces. The dynamics of the CML are defined by a map Φ that preserves the lattice structure;

$$\Phi x \;=\; (\Phi_\omega x)_{\omega\in\Omega} \tag{6.1}$$

where $\Phi_\omega : \mathcal{M} \to X_\omega$.

A popular approach in nonlinear dynamics considers the dynamics of the CML as a composition of two mappings: $\Phi = \mathcal{G} \circ \mathcal{F}$ where $(\mathcal{F}x)_\omega = f_\omega(x)$ is an independent action of local mappings $f_\omega : X_\omega \to X_\omega$, and $(\mathcal{G}x)_\omega = g_\omega(x)$ is an

interaction (Bunomovich, 1995; Kaneko, 1989b). The appeal to CMLs is due to their computational simplicity and their ability to exhibit a wide range of spatiotemporal phenomena.

For 1D-systems with nearest neighbour diffusive coupling, the state at time $t+1$ and site j, x_j^{t+1}, depends on the present state x_j^t and on the states of the two nearest neighbours, $[x_{j-1}^t, x_{j+1}^t]$. The nearest neighbour interaction destroys the low-dimensional deterministic behaviour at a given lattice site. Instead, a deterministic, infinite-dimensional system is formed. It has been found that where fundamental theory and approximate knowledge of the underlying state variables exists modelling of such systems essentially consists in parameter estimation (Ørstavik *et al.*, 1998). Analysis of spatiotemporal data from systems whose fundamental principles are unknown present a challenge. One promising approach is to apply methods derived in the case of low-dimensional systems. However, it is not guaranteed that genericity assumptions of Takens' theorem on the measurement functions on each site still hold. Therefore, whilst the approach is reasonable, the embedding obtained is not necessarily an embedding in the true spirit of the embedding theorems.

Proceeding with the ideas of Kaneko, the analysis and characterization of the highly complex behaviour of spatiotemporal systems is simplified by defining (for 1D-systems) a coupled map lattice as

$$x_j^{n+1} = (1-\epsilon)g(x_j^n) + \sum_{k=-l}^{r} \epsilon_k g(x_{j+k}^n) \tag{6.2}$$

where x_j^n is the state at site j and time n, and $g(x)$ is the nonlinear function describing the time evolution of the dynamics at each site. For conservation it is required that $\sum \epsilon_k = 1$. The coupled map lattice as defined couples $l \geq 0$ left neighbours and $r \geq 0$ right neighbours with coupling coefficients ϵ_k.

6.1.2 System Description

A 1D-CML system was investigated, with local dynamics given by the logistic map

$$f(x) \quad = \quad \alpha x(1-x) \tag{6.3}$$

with parameters $\alpha = 2.0$. It has been established that for this value of α chaotic behaviour is observed in the uncoupled logistic map(Abarbanel, 1996; Kantz and Schreiber, 1997). Assuming nearest neighbour diffusive coupling the CML system dynamics evolve according to

$$x_j^{n+1} \quad = \quad (1-\epsilon)f(x_j^n) + \frac{\epsilon}{2}\Big(f(x_{j-1}^n) + f(x_{j+1}^n)\Big) \tag{6.4}$$

6.1.3 Data Generation

Data used was generated by iterating the system in equation (6.4) for $\epsilon = 0.4$ using initial random conditions for a lattice of length $L = 100$, time $n = 150000$. Boundary conditions were assumed, that is, $x_{L=1}^n = x_{L=100}^n$. The first 140000 points were discarded to remove transient effects. The rest of the data was used to define training, validation, and test sets of length 2000, 1000, and 2000 points respectively for fitting a nonlinear model on the reconstructed state space.

Figure 6.1(a) is a time series plot showing the observed irregular behaviour. As in the ordinary logistic map, the unstable point is around 0.70, although for this system the instability is amplified. The trajectory traced at a single lattice site deviated significantly from the uncoupled case Figure 6.1(b). Assuming that the system dynamics are only temporal, it can be falsely concluded that the system is low-dimensional, the high-dimensional components being attributed to exogenous influences.

6.1.4 State space reconstruction

Reconstruction of state space is an ill-posed problem for coupled systems. Reconstruction of the spatiotemporal system was considered as follows: proceeding as

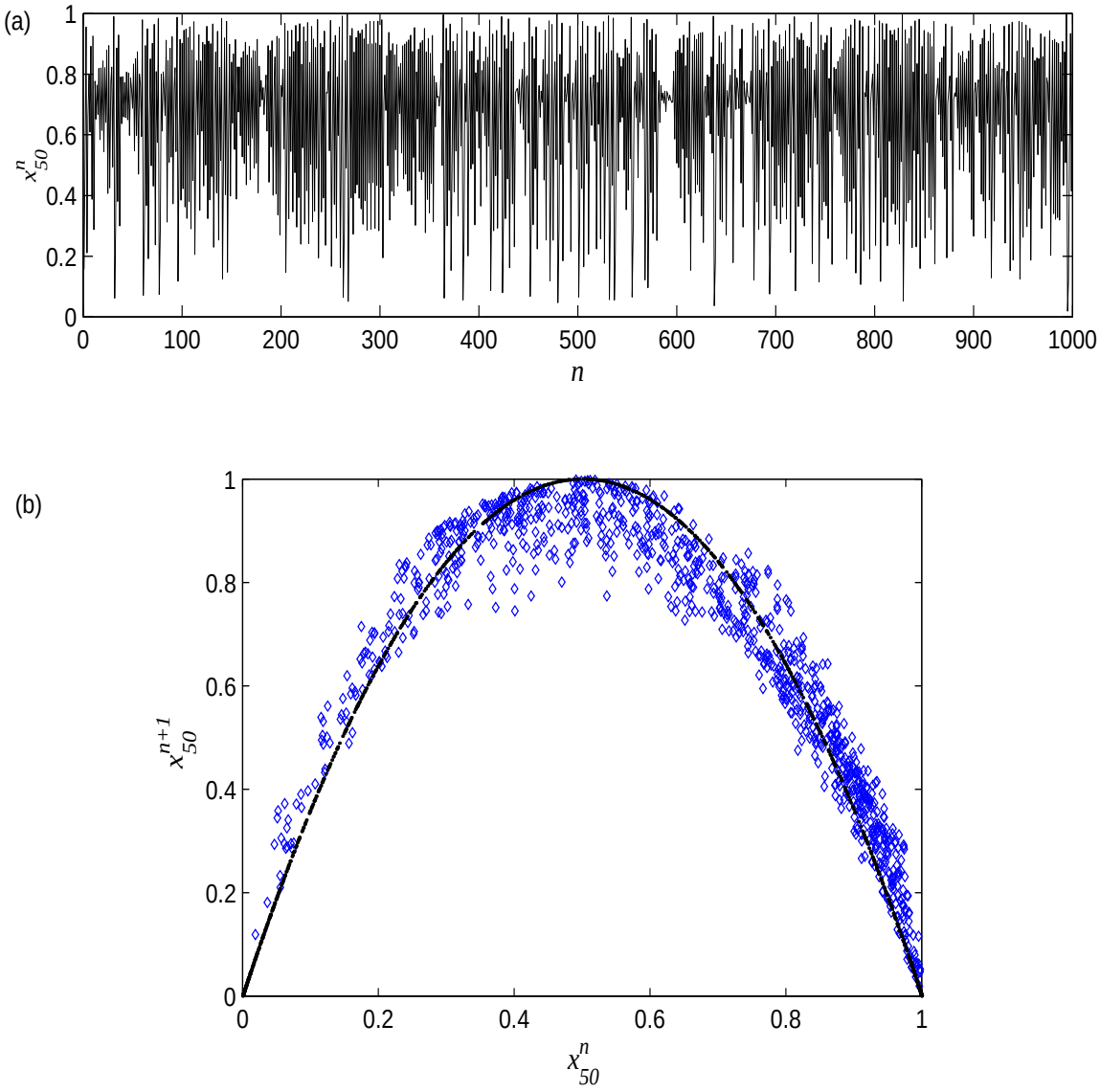


Figure 6.1: (a) Time series plot for 1000 successive points at a typical lattice site $j = 50$ of the 1D-coupled logistic map in equation (6.4) (b) Time delay reconstruction in a 2-dimensional embedding space for the data shown in Fig. 6.1(a). The effect of coupling on pattern dynamics is clearly revealed by the deviation from the behaviour exhibited by the uncoupled map superimposed on the plot.

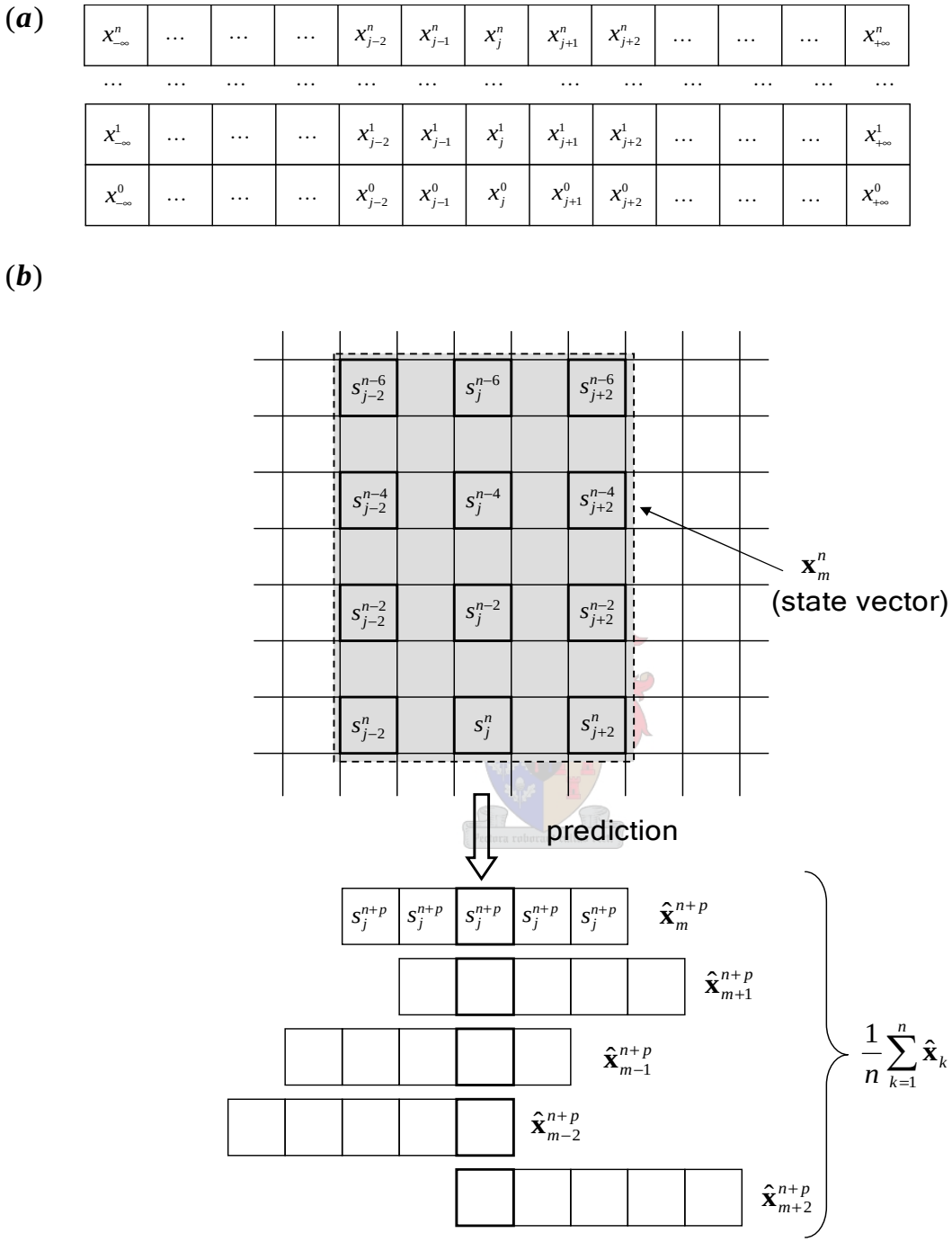


Figure 6.2: (a) A schematic coupled map lattice. (b) Reconstruction and prediction in spatiotemporal systems, $\{s\}_j^n$. In principle the average of the possible reconstructions must be used for prediction

Ørstavik *et al.* (1998), neighbours from only one side of reference lattice site $j = 50$ (referred to as x_0 subsequently) were included in the reconstruction. Furthermore, reconstructions that included neighbours from either side of the reference lattice site were additionally considered. Although the choice of the reference lattice site was arbitrary, similar results as reported later were expected for any choice of the reference site since boundary conditions were assumed in the experimental design. Rather than selecting an optimal embedding using either the false nearest neighbours or other similar constructs, it was decided to study the effect of varying both the temporal and spatial neighbours in the reconstruction, the rationale being that an embedding is not strictly defined for spatially extended systems. Also, since the local dynamics on the temporal evolution at each site was defined by a map, the natural choice for the time delay of $T = 1$ was used[1]

Proceeding as before, a nonlinear function f was defined on the dynamical behaviour of the reference site x_0^t using d_e temporal neighbours and d_s spatial neighbours taken either side of the reference site;

$$x_0^{t+1} = f \left\{ \begin{array}{c} (x^t_{+(d_s-1)}, x^{t-1}_{+(d_s-1)}, x^{t-2}_{+(d_s-1)}, \ldots, x^{t-(d_e-1)}_{+(d_s-1)} \\ x^t_{+(d_s-2)}, x^{t-1}_{+(d_s-2)}, x^{t-2}_{+(d_s-2)}, \ldots, x^{t-(d_e-1)}_{+(d_s-2)} \\ \ldots \\ x_0^t, x_0^{t-1}, x_0^{t-2}, \ldots, x_0^{t-(d_e-1)} \\ x_{-1}^t, x_{-1}^{t-1}, x_{-1}^{t-2}, \ldots, x_{-1}^{t-(d_e-1)} \\ \ldots \\ x^t_{-(d_s-1)}, x^{t-1}_{-(d_s-1)}, x^{t-2}_{-(d_s-1)}, \ldots, x^{t-(d_e-1)}_{-(d_s-1)} \end{array} \right\} \tag{6.5}$$

As shown in Figure 6.2(b), there are four possible predicted values in the case of a 1D-lattice depending on the reconstructed state vector used. A more accurate approach would have been to take the average these values for the predicted value. Sample results between this "overlap" approach and the simpler approach of using

[1]The autocorrelation function of maps have a first zero at $T = 1$.

only one value did not show significant differences. This was attributed this to the map studied and, therefore it was not found necessary to make broad generalization for other situations.

6.1.5 Characterization of spatiotemporal systems

Attractor invariants are important in discriminating time series generated by different dynamical systems. For example, one would need to know whether the system is evolving in a spatiotemporal chaotic regime, or is better explained by low-dimensional chaos. Generalization of known methods of analyzing low-dimensional systems to spatiotemporal systems is computationally demanding due to the very large number of degrees of freedom. In fact, dynamical invariants calculated for spatiotemporal systems are found to scale with subsystem size (Carretero-González, 1999). Instead, invariant densities are used, which are simply the estimated invariant measure divided by the system size. Because of this, algorithms for the estimation of the invariant quantities require the available data to increase exponentially with the dimension of the attractor (Ørstavik *et al.*, 2000), and therefore cannot be used in the same way as for low-dimensional system. Parekh *et al.* (1996) observed that the Lyapunov dimension and entropy increased linearly with the sub-system size while the Lyapunov dimension density rapidly saturated. Carretero-González (1999) and Ørstavik *et al.* (2000) introduced a new rescaling method for the estimation of the Lyapunov spectrum for spatially extended systems.

The CML system is a rich dynamical system that has been studied extensively in literature. Hence, it was assumed that the system as defined in equation 6.4 exhibited similar characteristic invariants as reported in literature. The task was then to define any optimal choice of d_e and d_s by fitting a nonlinear model to estimate f in equation 6.5. Ørstavik *et al.* (1998) followed more or less a similar approach although their work was limited to local linear modelling.

6.1.6 Modelling results

In Tables 6.1 and 6.2 are summary results obtained in the two reconstruction scenarios investigated, that is, 'one-sided' and 'two-sided' approaches respectively. In either case, the number of spatial and temporal neighbours was varied. Observing that the main objective was to investigate the effect of including spatial information versus only temporal information on a single site, the actual values of the mean square error performance criterion were normalized using the''best' model defined for the strictly temporal embedding case. Specifically, the model with temporal embedding dimension of 4 and an absolute $MSE = 0.0197$ was used as the reference model f_0. Hence, all results are relative to this case (and, therefore, with a relative value of 1.00, underlined in Table 6.1). Figures 6.3 and 6.4 are graphical plots of the results.

Table 6.1: Variation of mean square error and regression coefficient with embedding dimensions when only spatial neighbours from one side of x_0 are considered.

(d_e)	Spatial embedding dimension (d_s)									
	1		2		3		4		5	
	$\frac{mse_{f_0}}{mse_{f_i}}$	R^2	$\frac{mse_{f_0}}{mse_{f_i}}$	R^2	$\frac{mse_{f_0}}{mse_{f_i}}$	R^2	$\frac{mse_{f_0}}{mse_{f_i}}$	R^2	$\frac{mse_{f_0}}{mse_{f_i}}$	R^2
1	0.44	0.9609	0.70	0.9754	0.86	0.9807	0.84	0.9797	0.83	0.9772
2	0.61	0.9735	*2.43*	*0.9927*	*2.77*	*0.9937*	0.08	0.9389	0.08	0.8707
3	0.92	0.9763	*3.28*	*0.9947*	*1.59*	*0.9887*	*2.32*	*0.9923*	0.30	0.9511
4	<u>1.00</u>	0.9845	*1.95*	*0.9913*	*1.61*	*0.9892*	*1.61*	*0.9893*	0.67	0.9756
5	0.82	0.9867	*2.37*	*0.9926*	0.80	0.9811	*1.11*	*0.9871*	0.28	0.9480

Table 6.2: Variation of mean square error and regression coefficient with embedding dimensions. Here $d_{s\pm}$ means this number of spatial dimensions either side of x_0.

d_e	Spatial embedding dimension ($d_{s\pm}$)									
	1		2		3		4		5	
	$\frac{mse_{f_0}}{mse_{f_i}}$	R^2	$\frac{mse_{f_0}}{mse_{f_i}}$	R^2	$\frac{mse_{f_0}}{mse_{f_i}}$	R^2	$\frac{mse_{f_0}}{mse_{f_i}}$	R^2	$\frac{mse_{f_0}}{mse_{f_i}}$	R^2
1	*7.1E+07*	*1*	*1.29E+06*	*1*	*2.56E+05*	*1.00*	7.54E+01	0.9998	3.87E+03	1.00
2	*2.7E+05*	*1*	*1.45E+03*	*1*	*1.72E+01*	*1.00*	3.67E+01	0.9998	8.46E-01	0.9889
3	*7.1E+05*	*1*	*4.97E+03*	*1*	*2.66E+00*	*0.9972*	5.38E-01	0.9827	1.77E+02	1.00
4	*8.7E+04*	*1*	*1.95E+03*	*1*	*9.14E+01*	*0.9999*	5.01E-01	0.9797	6.92E+01	0.9999
5	*1.9E+02*	*1*	*3.67E+03*	*1*	*2.72E+00*	*0.9961*	1.11E+01	0.9988	1.75E-01	0.9238

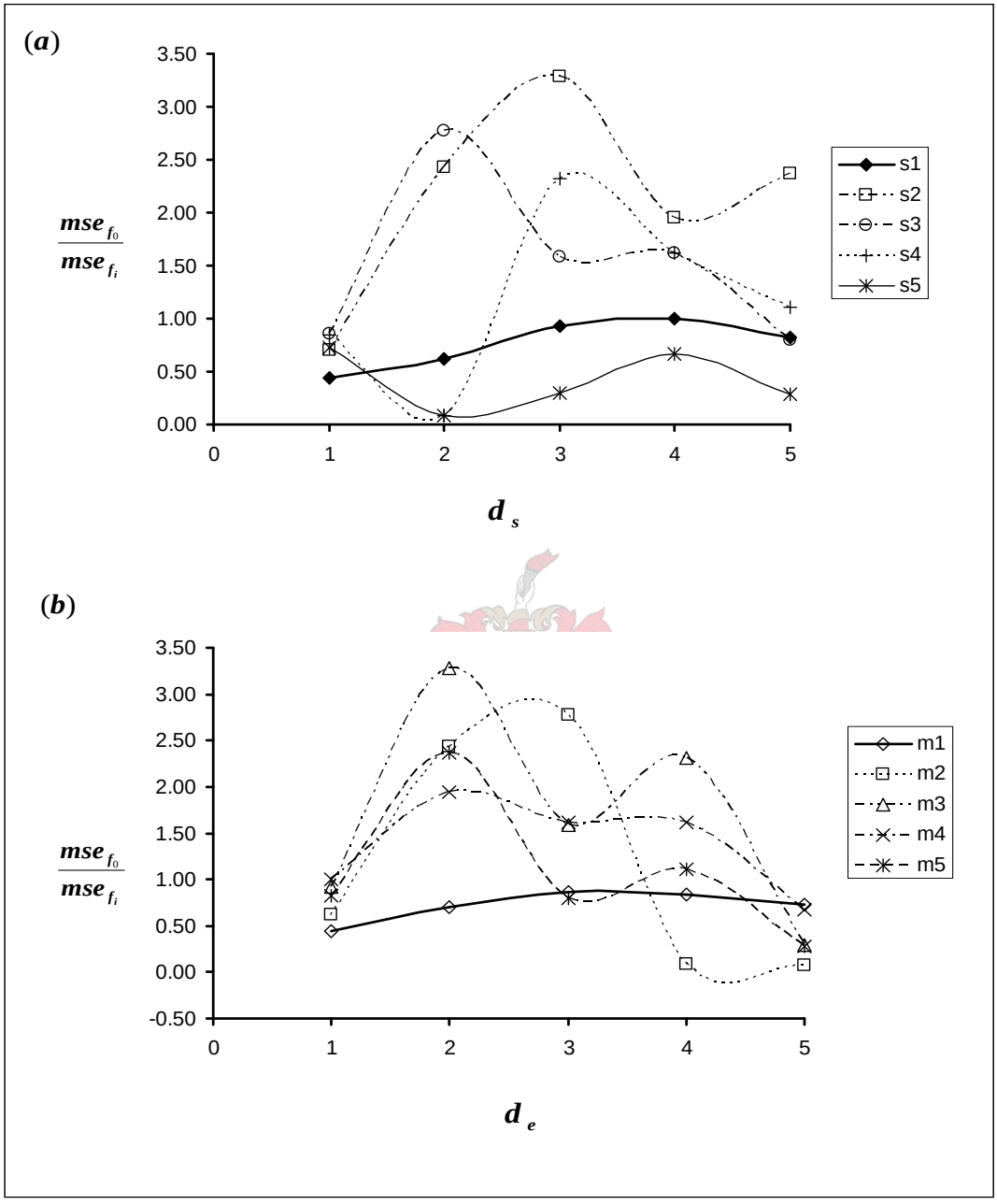


Figure 6.3: (a) Effect of varying the temporal neighbours on model performance for different spatial reconstruction dimensions (b) Effect of varying the spatial neighbours taken from one side of the reference lattice site on model performance for different temporal embeddings

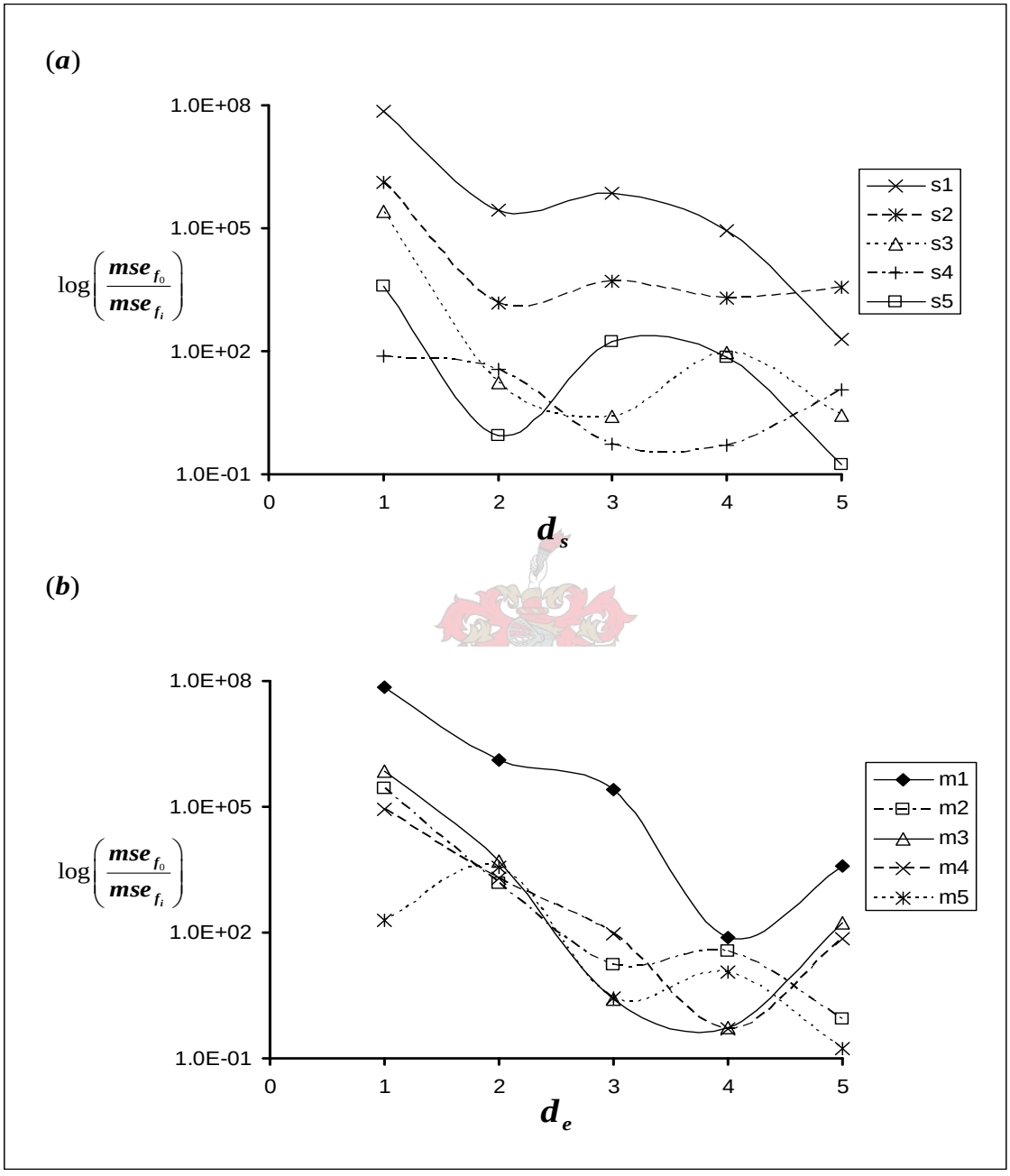


Figure 6.4: (a) Effect of varying the temporal neighbours on model performance for different spatial reconstruction dimensions (b) Effect of varying the spatial neighbours taken from both sides of the reference lattice site on model performance for different temporal embeddings. The difference with Figure 6.3, is that here the effective spatial dimension d_s is $\frac{3}{2}\times$ the same d_s in the other plot

6.1.7 Discussion

For models using reconstructions that included neighbours from one side of the reference site the following were observed:

- Inclusion of spatial information generally reduce the mean square error of predictive models but only for certain choices of the spatial neighbours, in particular, for $1 \leq d_s \leq 3$. For $d_s = 4$, the performance is only better for $d_e > 2$.
- For $d_s \geq 5$, inclusion of spatial information degrades model performance.
- Inclusion of more temporal information from further in the past gave better models for specific spatial dimensions, $d_s \leq 4$. In general, different choices of the spatial neighbours gave an optimal performance at different d_e. No particular pattern was discernible that could indicate how this choice depended on the number of spatial neighbours.
- The optimal model was defined for $d_e = 3, d_s = 2$.

In Table 6.1 the reconstruction parametric space indicating the zone of improved model performance than the reference model f_0 is indicated with a smaller, slanted font.

In the case of models using reconstructions that included bi-directional spatial information, the following were observed:

- As would be expected, inclusion of spatial information from both sides of the reference site gave a phenomenal decrease in the cost function, with the best model defined for $d_{s\pm} = 1, d_e = 1$.
- A general decreasing trend in model performance with increasing temporal dimension for all cases.

- Increase of spatial neighbours also had a negative effect on the resulting model. In fact, for all $d_e < 5$, corresponding reconstructions gave optimal improved models for $d_{s\pm} = 1$. For $d_{s\pm} \leq 2$ all reconstructions had a mean square error $MSE \gg MSE_{f_0}$. For other choices of $d_{s\pm}$ model performance became a function of the temporal neighbours used.

- Near perfect model fits were obtained for most of the reconstructions with regression coefficient values equal to unity, that is $R^2 = 1$.

Table 6.2 shows the reconstruction parametric space indicating the zone of improved model performance with respect to the reference model f_0. The region is shown with a smaller slanted font.

6.2 Concluding Remarks

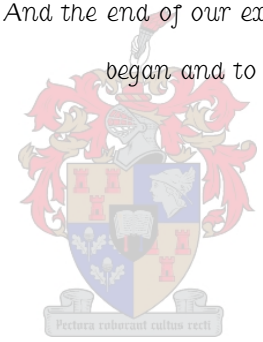
From the results presented it can be concluded that use of spatial information in addition to temporal delays is comparatively superior to either pure spatial reconstruction or temporal reconstruction only. However, the advantages of using spatial and temporal information is restricted to narrow ranges. In particular, it was observed that using only the nearest neighbours is optimal regardless of the choice of the temporal delays. However, the challenges of defining a proper model for the local dynamics, the flow of information, and the strength of coupling for practical systems is still largely unresolved. It is necessary to validate the results obtained by using a controlled experiment where measurements are taken from different sites. These measurements could be from image data taken on pattern forming system for example.

Chapter 7

Conclusions & Recommendations

"We must not cease from exploration. And the end of our exploring will be to arrive where we began and to know the place for the first time"

– TS Elliot



7.1 Conclusions

Below are listed the main conclusions from the results reported in this work:

- As expected, models obtained using multivariate extension of the embedding theorems were generally found to exhibit better predictability than corresponding models from scalar embedding. In the modeling of the coupled CSTR system it was found that the choice of channels used in the multivariate reconstruction affected the performance of resulting model. Certain combinations gave relatively better models than others. The improved performance of multivariate models was related to the decrease in uncertainty over the interpoint distribution of points in the reconstructed state space. Furthermore, local linear models indicated that use of multivariate embedding captured the underlying determinism over a broader region in phase space

compared to univariate schemes.

- Better and consistent predictive models were obtained in non-uniform embedding strategies compared to univariate reconstructions. This was attributed to the capture of dynamics over multiple timescales. Uniform embedding use a single timescale defined by the time lag between successive components in the embedded vectors. It was also observed that such models were less dependent on the choice of components used in reconstructing the attractor. Furthermore, the non-uniform embedding strategy facilitated for simultaneous optimization of the both the reconstruction and parameterized model from a given model class.

- Use of multivariate signals gave a seemingly more robust nonlinearity test for observed signals using constrained-realization surrogate analysis. Preservation of both autocorrelations and cross-correlations existing among the observed signals decreased the risk of false detection of nonlinearity in data. However, the surrogate generating algorithms could not always constrain the surrogates to mimic the linear properties of multiple channels as was observed by in certain surrogates that had wildly nonsensical or erroneous entropy values.

- The correlation dimension estimate as a function of viewing scale, $d_c(\varepsilon_0)$ provided a better discriminating statistic than the correlation dimension estimate $d_c(m)$ and entropy $K(m)$as functions of the embedding dimension in nonlinearity tests. This was particularly true for multivariate surrogate analysis. Although the maximal Lyapunov exponent is often used to identify chaotic behaviour in physical systems, it was observed that it was difficult to get correct and consistent estimates because of sparse and noisy data using current algorithms.

- In the modeling of a flotation process, least-squares support vector machines

performed better than multilayer perceptron networks. However, multilayer perceptron gave better models in modeling the coupled CSTR system. It was also found that the performance of the MLP degraded with increasing prediction time step. MLPs performed worse than LS-SVMs at longer one-step ahead prediction time steps. From the different performances of the two models classes in modeling with the flotation process and coupled CSTR, no general conclusions with regard to the comparative merits of the two approaches could be made except that when autocorrelations of the variables were weak and the time series short, such as the flotation data, LS-SVMs performed better than MLPs. Consistent results from the simulation of a trained LS-SVM were obtained, which was not the case for MLPs.

The computational cost of selecting the optimal hyperparameters for the LS-SVMs was significantly higher than the cost incurred in optimizing the MLP network parameters. This was despite the fact that only a subset of the data was used in selecting the hyperparameters of LS-SVMs whereas the entire training data set was used in adjusting MLP parameters. However, the simulation time of both learning algorithms were comparable for the data used.

- In the case of a coupled CSTR system use of either independent component analysis (ICA) or principal component analysis (PCA) with whitening did not improve the predictive capabilities of the resulting models over models that used only PCA reduction. In particular, the performance of MLPs degraded with increasing complexity of the separation algorithm whereas LS-SVM models were insensitive to the separation method used.

- In modeling time series data from a coupled map lattice it was found that including spatial information in the reconstruction resulted in a remarkable improvement in model predictive potential. Also, inclusion of all nearest neigh-

bours gave the best one-step ahead predictive models. Hence spatiotemporal time series analysis offer potentially superior predictive modeling capabilities compared to models based on individual time series.

7.2 Recommendations for Future Work

Admittedly, the investigation barely scratched the surface of issues and ambiguities encountered in multivariate time series. However, the results highlight areas that, if resolved, could offer better practical opportunities in extending the tools and techniques in modeling and control of process operations. These research opportunities are discussed briefly in the following paragraphs.

In some cases, multivariate time series analysis offer potential advantages over scalar embeddings. The practical extensions of the embedding theorems used in this report were *ad hoc* and pre-supposed or pre-empted a generalized embedding theorem. However, the rigorous mathematical proofs of embedding theorems exist only for scalar time series. It does not necessarily follow that the generic assumption of these embedding theorems still hold when dealing with multivariate time series. Further work within theoretical nonlinear dynamics is needed to address this aspect .

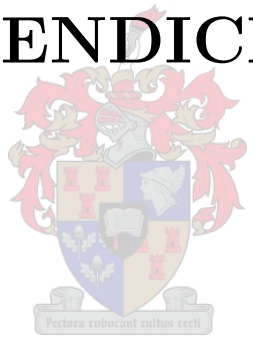
The practical applications of nonlinear modeling including methods based on chaos theory are not yet well-established in the process industries. This is a result of many factors including the fact that there still exists many as yet unresolved theoretical issues, and the complexity of the methods involved. To realize the full benefit and implications of especially multivariate time series nonlinear system identification and control design theory should be integrated. Specifically, since modern control systems are model-based, controller design must be able to identify when the parameterized model fails in approximating process behaviour because of

drift in the process parameters, for example. It is important for the controller to compensate for the instabilities that real process systems present by including such knowledge in the integrated model-controller-process setup.

In relation to the two nonlinear model classes used here, a definitive superiority of one method over the other could only have been established in terms of some criterion based on, for example, minimum description length or VC-dimension. The theoretical concepts of how to proceed are fairly well-established. However, there is yet little in terms of algorithmic implementation. Such a clear superiority of one method over the other may never be established. Rather, a framework which provides for guidelines in choosing the suitable method to use for given circumstances could be developed.

Variable embedding strategies based on the concept of non-uniform embedding offer a promising route in building models that describe the long-term behaviour of nonlinear processes. These strategies have been used in, for example, radial basis networks. Extending the approach to other model classes is needed. It is not expected that the extensions will be generalization of the current implementations because of intrinsic differences of various model classes.

APPENDICES



Appendix A

Phase Space Reconstruction

The equations of motion of dynamical systems are defined in terms of first order differential equations acting on a finite-dimensional phase space, $\mathbb{R}^d$.

$$\dot{\mathbf{x}} \;=\; \mathbf{f}(t, \mathbf{x}(t)), t \in \mathbb{R} \tag{A.1}$$

where the vector $\mathbf{x}$ consisting of d *independent* components (or effective degrees of freeedom) specifies the state at time t. The effective degrees of freedom are the number of initial conditions required to specify the dynamical system. The phase space is a mathematical space with orthogonal co-ordinate directions representing each of the variables needed to specify the instantaneous state of the system. The phase space is not necessarily equal to the spatial dimension of the dynamical system, although the spatial dimension sets the upper limit on the values the phase space can take. Given only a time series, the embedding theorems guarantee that the reconstructed phase space of dimension d_e is similar to the underlying phase space in $\mathbb{R}^d$ except for some co-ordinate change. Figure A.1 below shows graphically a typical reconstruction of the phase space using data generated from an iterated

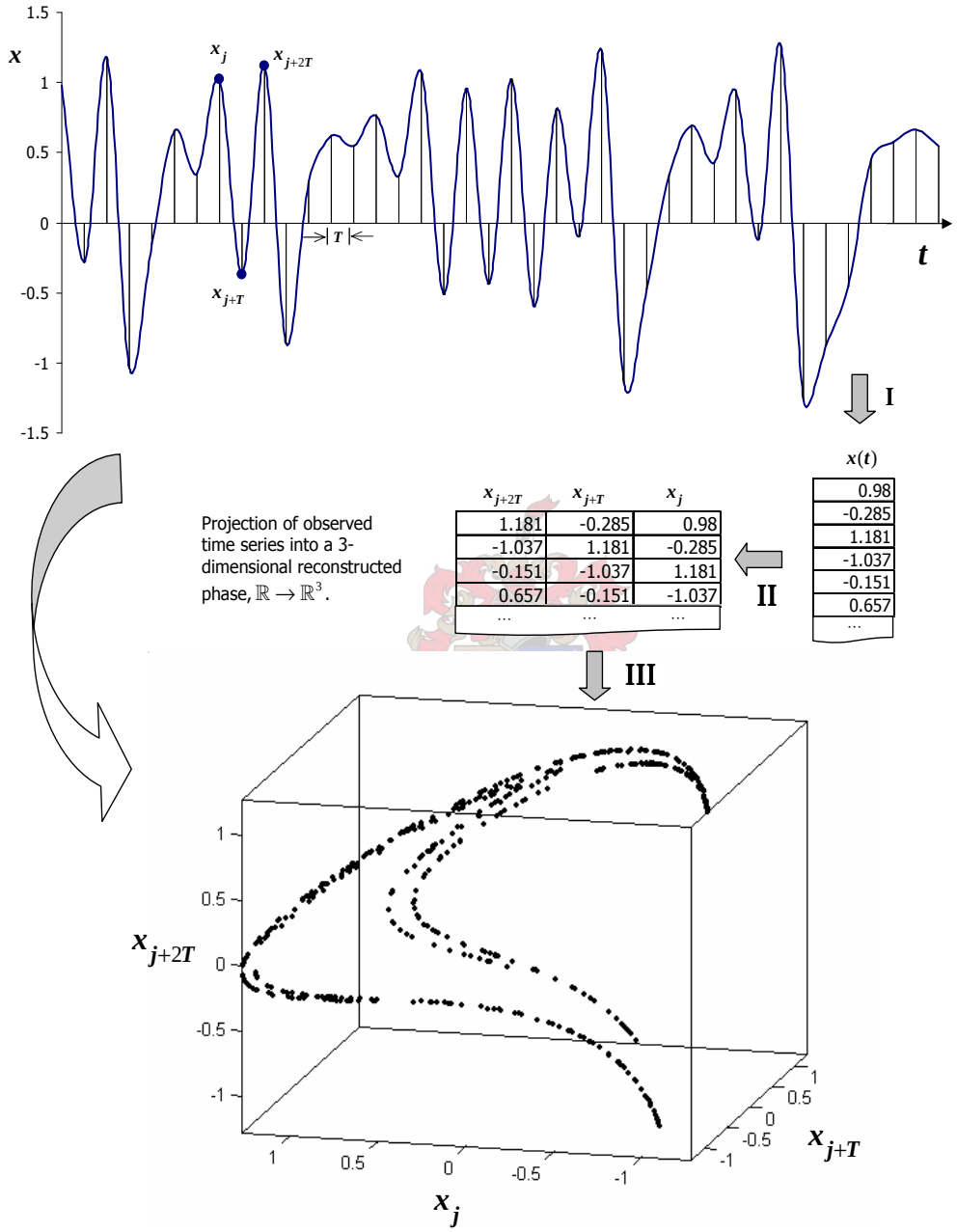


Figure A.1: An illustration of the phase space reconstruction process applied to data from a Henon map. I. Uniformly sampled, time-ordered values from a signal observed on the physical system. II. Reconstruction of the phase space from vectorized independent components created from the time series. III. The unfolded reconstructed attractor viewed in the embedded space

Hénon map in the chaotic regime;

$$\begin{aligned} x(n) &= y(n-1) - 2.8x(n-1)^2 + 1 \\ y(n) &= 0.3x(n-1) \end{aligned} \tag{A.2}$$

Appendix B

Sampling Theory of Correlation

N pairs of values (x, y) of two random variables X and Y constitute a bivariate population, which is assumed to be a bivariate normal distribution. The theoretical population coefficient of correlation, ρ, is estimated by the sample correlation coefficient r, or $\hat{\rho}$, defined as

$$\begin{aligned}\hat{\rho} &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \\ &= \frac{n\sum_i^N x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}}\sqrt{n\sum x_i^2 - (\sum x_i)^2}\end{aligned}$$

Tests of siginificance or hypotheses concerning various values of ρ require knowledge of the sampling distribution of $\hat{\rho}$. For $\rho = 0$ this distribution is symmmetrical, a statistic involving Student's t distribution can be used. For $\rho \neq 0$, the distribution is skewed. A statistic with an approximately normal distribution is obtained using Fisher's Z transformation:

$$Z = \frac{1}{2}\log_e\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right)$$

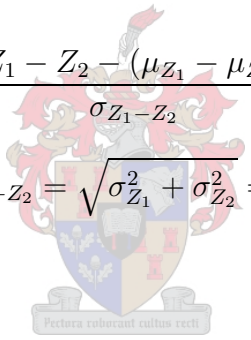
Z is approximately normally distributed with mean and standard deviation given by;

$$\begin{aligned} \mu_Z &= \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right) \\ \sigma_z &= \frac{1}{\sqrt{N-3}} \end{aligned}$$

The significance of a difference between correlation coefficients , ρ_1 and ρ_2, drawn from samples of sizes N_1 and N_2 respectively, can be computed using Fischer's Z transformation and the fact that the test statistic

$$z = \frac{Z_1 - Z_2 - (\mu_{Z_1} - \mu_{Z_2})}{\sigma_{Z_1 - Z_2}}$$

is normally distributed, where $\sigma_{Z_1-Z_2} = \sqrt{\sigma_{Z_1}^2 + \sigma_{Z_2}^2} = \sqrt{\frac{1}{N_1-3} + \frac{1}{N_2-3}}$.



Appendix C

The Method of Surrogate Data

The implementation of the method of surrogate data for the detection of nonlinearity is illustrated for the null hypothesis that the observed time series is a nonlinear static transformation of a linear stochastic process. The original time series is first rescaled to have a Gaussian distribution. A surrogate time series which has the same Fourier spectrum as the rescaled original is generated by phase randomisation. This surrogate is then rescaled to have the same values as the original time series, resulting in a surrogate time series with a probability distribution similar to the original time series. This method is also called ***A*lgorithm II**. Details can be found in Theiler *et al.* (1992) and Schreiber and Schmitz (2000).

1. Sort the original time series, $x(t)$, $Sx(k)$, $\quad k = 1, \ldots, N$

$$\begin{aligned} x(t) &= [1.0051; -0.7397; 1.282; -1.7777; \\ &\quad -0.55518; 0.55959; 0.13621; 1.0677; \\ &\quad -0.85329; 1.1971] \\ Sx(k) &= [-1.7777; -0.85329; -0.7397; -0.55518; \\ &\quad 0.13621; 0.55959; 1.0051; 1.0677 \\ &\quad 1.1971; 1.282] \end{aligned}$$

2. Make ranked time series $Rx(t)$, defined to satisfy $Sx[Rx(t)] = x(t)$.

$$Rx(t) = [4; 9; 2; 5; 7; 6; 1; 8; 10; 3]$$

3. $Sx(k)$ is a monotonic function with a well-defined inverse; so $Rx(t) = Sx^{-1}(t)$ is a static rescaling of $x(t)$.

4. Create a random Gaussian data set $g(t)$, $\quad t = 1, \ldots, n$.

$$\begin{aligned} g(t) &= [-0.70543; -1.44; -0.10684; -0.7056; \\ &\quad -1.0265; 0.29297; 0.32556; -0.75657; \\ &\quad 0.82143; -1.1129] \end{aligned}$$

5. Sort the Gaussian random numbers $Sg(k)$, $\quad k = 1, \ldots, n$.

$$\begin{aligned} Sg(k) &= [-1.44; -1.1129; -1.0265; -0.75657; \\ &\quad -0.7056; -0.70543; -0.10684; 0.29297; \\ &\quad 0.32556; 0.82143;] \end{aligned}$$

6. Define the new time series: $y(t) = Sg[Rx(t)]$.

$$\begin{aligned} y(t) &= [-0.10684; -1.0265; 0.82143; -1.44; \\ &\quad -0.75657; -0.70543; -0.7056; 0.29297; \\ &\quad -1.1129; 0.32556] \end{aligned}$$

$y(t)$ is a static rescaling of $x(t)$ with the property that the amplitude distribution is Gaussian.

7. Compute the discrete Fourier transform $y(f)$ of the Gaussian time series $y(t)$; $y(f) = \mathcal{F}(y(t)) = \sum_{n=0}^{N-1} y(t_n) e^{2\pi f n \Delta t} = A(f) e^{i\phi(f)}$, where $A(f)$ is the amplitude and $\phi(f)$ is the phase.

$$\begin{aligned} y(f) &= [-4.7395; 0.8944 - 0.077568i; -1.5727 + 0.93467i; -1.0089 - 0.42716i; \\ &\quad 3.5762 - 0.8958i; 3.5762 + 0.8958i; -1.0089 + 0.42716i; \\ &\quad -1.5727 - 0.93467i; 0.8944 + 0.077568i] \end{aligned}$$

8. Randomize the phases by rotating ϕ at each frequency f by an independent random variable φ chosen uniformly in the range $[0, \ldots, 2\pi]$.

$$y'(f) = A(f) e^{i[\phi(f) + \varphi(f)]}$$

9. Symmetrize the phases: $\mathrm{Re}\{y''(f)\} = \mathrm{Re}\{(y'(t) + y'(n+1-f))/2\}$, $\mathrm{Imag}\{y''(f)\} = \mathrm{Imag}\{(y'(f) + y'(n+1-f))/2\}$;

$$\begin{aligned} y''(f) &= [4.7395; -0.88726 + 0.13686i; 0.86648 - 1.6113i; \\ &\quad 0.74666 - 0.80181i; 3.5762 + 0.8958i; 0.74666 + 0.80181i; \\ &\quad 0.86648 + 1.6113i; -0.88726 - 0.13686i] \end{aligned}$$

10. Do a Fourier Transform inverse: $y'(t) = \mathcal{F}^{-1}\{y''(f)\} = \mathcal{F}^{-1}\{y'(f)e^{i\varphi(f)}\}$

$$\begin{aligned} y'(t) &= [0.97674; -0.25129; 0.6252; 0.56128; \\ &\quad 0.42351; 1.033; 0.14331; 0.40375; \\ &\quad 0.57432; 0.2497] \end{aligned}$$

$y'(t)$ is the surrogate data of $y(t)$.

11. Rank $y'(t)$ to form $Ry'(t)$.

$$Ry'(t) = [9; 1; 8; 6; 5; 10; 2; 4; 7; 3]$$

12. The surrogate time series is then given by $x'(t) = Sx[Ry'(t)]$.

$$\begin{aligned} x'(t) &= [1.1971; -1.7777; 1.0677; 0.55959; \\ &\quad 0.13621; 1.282; -0.85329; -0.55518; \\ &\quad 1.0051; -0.7397] \end{aligned}$$

The surrogate time series $x'(t)$ is just a shuffling of the observed time series $x(t)$, and therefore their amplitude distribution. Further if we define G as the transformation from the amplitude distribution of x to a Gaussian amplitude distribution, then we have the property that $G(x)$ has the same Fourier power spectrum (and hence, the same autocorrelation) as $G(x')$. Note $G = h^{-1}$ where h is the measurement function of the null hypothesis.

Figure C.1 illustrates is a graphical depiction of the surrogate generation process for a time series generated using a henon map.

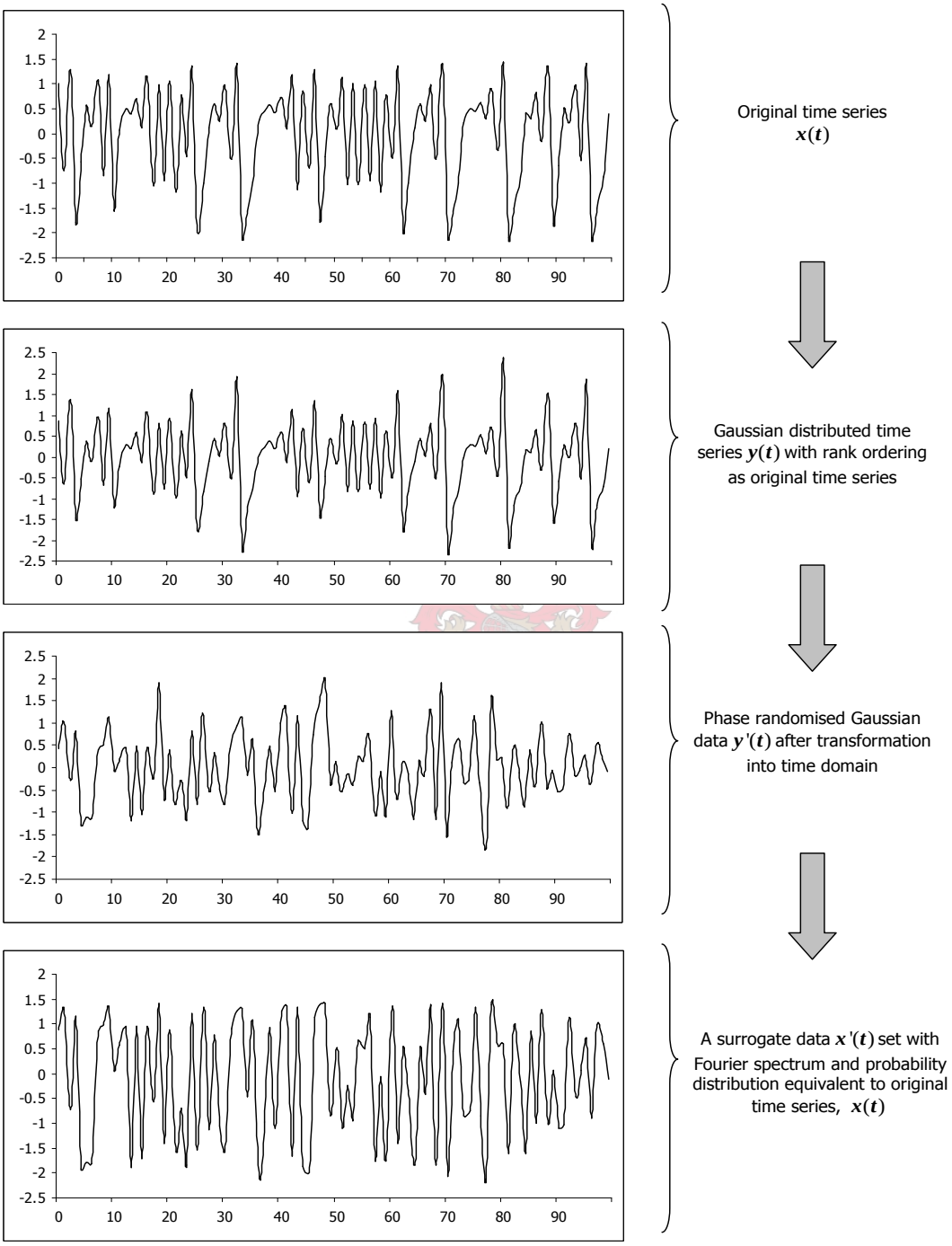


Figure C.1: A graphical illustration of transformations involved in Fourier-based surrogate data generation

Appendix D

The Grassberger-Procaccia Algorithm for D_2 Estimation

The correlation dimension d_c is the most widely used dimension estimate for attractors because of computational simplicity of the algorithm of Grassberger and Procaccia (1983), on which most implementations are based on. The GP algorithm uses the correlation sum $C(r)$ to determine the interpoint distribution of points in the phase space (see Chapter 2). The correlation sum scales with the hypersphere diameter r defined around a point on the trajectory according to a power law of the form;

$$C(r) \quad \propto \quad r^{d_c} \tag{D.1}$$

An examination of the attractor for many different hypersphere radii, d_c is obtained from the slope of the scaling region of a $\log(r) - log[C(r)]$ plot, Figure D.1. $C(r) \approx 1$ for large radii, as all points on the attractor are contained in the hypesphere. The plot tapers off at large radii as increases in further increases in radius result in increasingly smaller changes in the correlation sum. Similarly, at smaller radii fluctuations in the number of points contributing to $C(r)$ result in fluctuations in the plot.

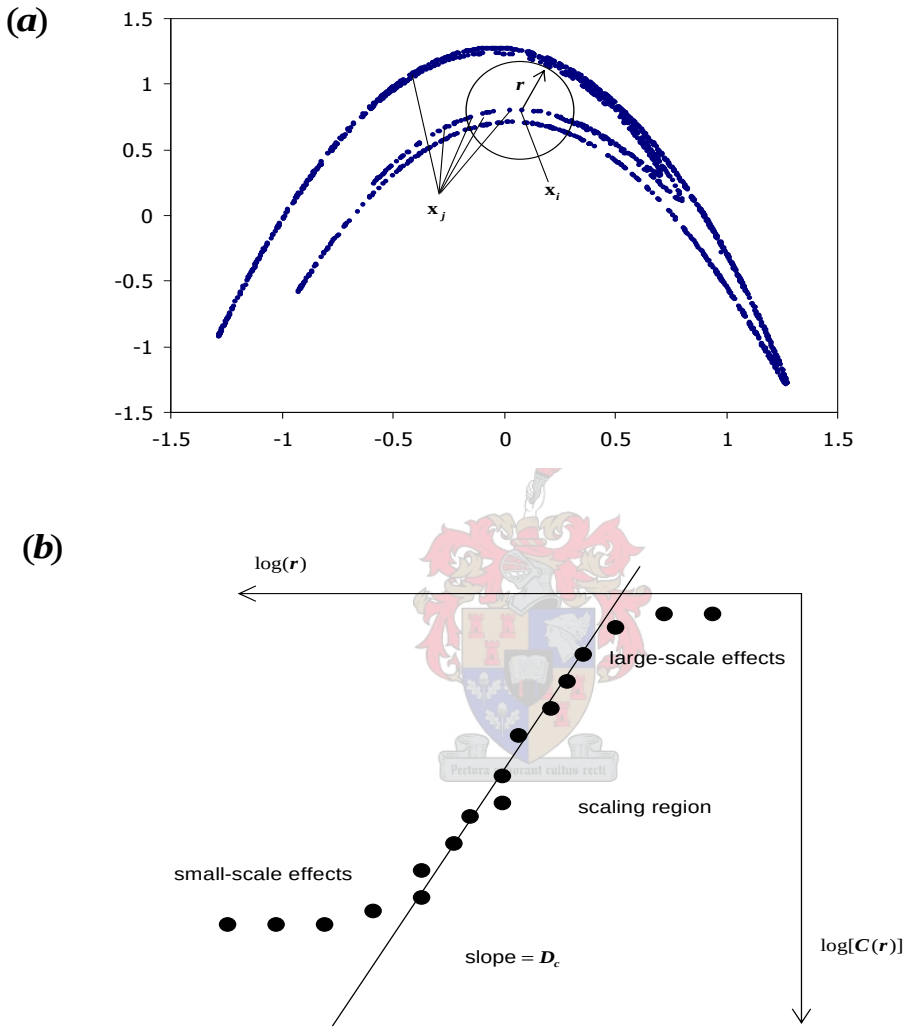


Figure D.1: Estimating the correlation using Grassberger-Procaccia approach. (a) Calculating probability of interpoint distribution on attractor. (b) The $\log(r) - \log[C(r)]$ plot.

Glossary

Attractor – The subset of phase space in which the trajectories of a dynamical system asymptotically collapse onto with time.

Chaos – Nonlinear behaviour visually indistinguishable from a stochastic signal generated by a physical system whose dynamics are governed by deterministic laws.

Embedding – A one-to-one mapping function from the attractor into a reconstructed phase space that preserves differential information.

Entropy – An invariant nonlinear statistical quantity which gives the information properties of an attractor.

Ergodicity – A property shared by nonlinear statistical quantities used to characterize attractors. An *ergodic* measure is an indecomposable quantity that is invariant under the action of the dynamic system. In simple terms ergodicity says that a time average is equal to an "ensemble" or space average.

Generalized Dimensions – Invariant nonlinear statistical quantities that characterize the geometrical properties of an attractor.

Lyapunov Exponents – Invariant nonlinear statistical measures that characterize the sensitivity to initial conditions a dynamical system is. The total number of Lyapunov exponents equals the state space dimensions. For chaotic systems at least one of the exponents is strictly positive.

Machine Learning – Extraction of a functional relationship mapping an input space to an output space using algorithmic formulations such as neural networks

and support vector machines. When properly learned, it is possible to use the "machine" for prediction when presented with future inputs.

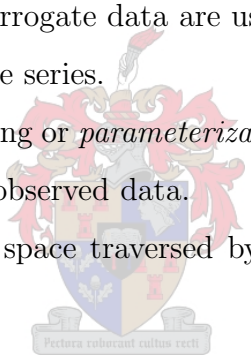
Phase/State Space – A multidimensional vector space in which a dynamical system evolves. Typically, the number of dimensions of this phase space equal the number of degrees of freedom the system has.

State Space Reconstruction – Inference of dynamical properties of a physical system using observed time series.

Surrogate Data – Artificial or computer generated data that have the same linear properties as the observed data. In particular, surrogate data have the same probability distribution and Fourier spectrum (strictly speaking, autocorrelation function) as the observed data. Surrogate data are useful in bootstrap approaches for detection of nonlinearity in time series.

System Identification – The fitting or *parameterization* of a mathematical model to a dynamical system using only observed data.

Trajectory – The path in phase space traversed by the dynamical system with time.



Bibliography

Abarbanel, H.D.I, Analysis of observed chaotic data, *Springer*, Berlin, (1996).

Abarbanel, H.D.I., Brown, R., Sidorowich, J.J., and Tsimring, L.S., The analysis of observed chaotic data in physical systems, *Rev. Mod. Phys.*, **65**(4), (1993), pp. 1331 – 1391.

Abarbanel, H., Carroll, T., Pecora, L., Sidorowich, J., and Tsimring, Predicting physical variables in time–delay embedding, *Phys. Rev. E*, **49**, (1994), pp. 1840–1853.

Abashar, M.E.E. and Judd, M.R., Synchronization of chaotic nonlinear oscillators: study of two coupled CSTRs, *Chem. Eng. Sci.*, **53**(21), (1998), pp. 3741 – 3750.

Alhumaizi, K. and Aris R., Chaos in a simple two-phase reactor, *Chaos, Solitons & Fractals*, **4**, 1994, pp. 1985–2014.

Ajbar, A., Stabilization of chaotic behavior in a two-phase autocatalytic reactor, *Chaos, Solitons and Fractals*, **12**, (2001), pp. 903–918.

Barnard, J.P., Empirical state space modelling with applications in online diagnosis of multivariate dynamic systems, *PhD Thesis*, University of Stellenbosch, (1999).

Barnard, J.P., Aldrich, C., and Gerber, M., Embedding of multidimensional time-dependent observations, *Phys. Rev. E*, **64**, (2001).

Boccaletti, S., Kurths, J., Osipovd, G., Valladares, D.L., and Zhou, C.S., The synchronization of chaotic systems, *Phys. Rep.*, **366**, (2002), pp. 1–101.

Box, G.E. and Jenkins, G.M., Time Series Analysis, *Holden-Day*, San Francisco, (1976).

Bunimovich, L.A., Coupled map lattices: one step forward and two steps back, *Physica D*, **86**, (1995), pp. 248 – 255.

Cao, L., Practical method for determining the minimum embedding dimension of a scalar time series, *Physica D.*, **110**, (1997), pp. 43 – 50.

Cao, L., Mees, A., and Judd, K., Dynamics from multivariate time series., *Physica D*, **121**, (1998), pp. 75–88.

Casdagli, M., Chaos and deterministic versus stochastic nonlinear modeling, *J. Roy. Stat. Soc. B*, **54**, (1991), pp. 303–238.

Carretero-González, R., Ørstavik, S., Huke, J., Broomhead, D.S., and Stark, J., Scaling and interleaving of sub-system Lyapunov exponents for spatio-temporal systems, *Chaos*, **9**(2), (1999).

Casdagli, M., A Dynamical Systems Approach to Modeling Input–Output Systems, in: Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, Proc. Vol. XII, *Addison-Wesley*, eds. M. Casdagli and S. Eubank, (1992), pp. 265–281.

Casdagli, M., Eubank, S, Farmer, J.D., and Gibson, J., State space reconstruction in the presence of noise, *Physica D*, **51**, (1991), pp. 52–98.

Cottis, R.A., Al-Awadhi, M.A.A., Al-Mazeedi, and Turgoose, S., Measures for the detection of localized corrosion with electrochemical noise, *Electro. Acta.*, **46**, (2001), pp. 3665–3674.

Cristianini, N. and Shawe-Taylor, J., An Introduction to Support Vector Machines, Cambridge University Press, (2000).

Cross, M.C., and Hohenberg, P.C., Pattern formation outside of equilibrium, *Rev. Mod. Phys.*, **65**(3), (1993), pp. 851–1123.

Crutchfield, J.P., Knowledge and Meaning ... Chaos and Complexity, in Modeling Complex Phenomena, eds. L. Lam and V. Naroditsky, Springer, Berlin, (1992), pp. 66–101.

Diks, C., Takens, F., and DeGoede, J., Spatio-temporal chaos: A solvable model, *Physica D*, **104**, (1997), pp. 269–285.

Duan K., Keerthi, S.S., Poo A.N., Evaluation of Simple Performance Measures for Tuning SVM Hyperparameters, *preprint*, Available at: `http://guppy.mpe.nus.edu.sg/~mpessk/papers/comparison.pdf`, (2002).

Eckmann, J.P., and Ruelle, D., Ergodic theory of chaos and strange attractors, *Rev. Mod. Phys.*, **57**(3), (1985), pp. 617–656.

Elnashaie, S.S.E.H., Abashar, M.E., and Teymour, F.A., Chaotic behaviour of fluidized-bed catalytic reactors with consecutive exothermic chemical reactions, *Chem. Eng. Sci.*, **50**, (1995), pp. 49–67.

Eneva, E., Kumaraswamy, K., and Watteucci, W., A Study in Fractal Dimension and Dimensionality Reduction, Technical Report, CALD, Availabel at: `www.cs.cmu.edu/~eneva/wekkem.pswww`, (2002).

Finney, C.E.A., Kennel, M.B., Daw, C.S., and Halow, J.S., Nonlinear time-series diagnostics of fluidization quality, *presented at the Annual AIChE*, (1996); Available at: `http://www-chaos.engr.utk.edu/pap/crg-aiche1996-slides.pdf`, (2002)

Fraser, A.M. and Swinney, H., Independent coordinates for strange attractors from mutual information, *Phys. Rev. A*, **33**, (1986), pp. 1134–1140.

Grassberger, P. and Procaccia, I., Characterization of strange attractors, *Phys. Rev. Let.*, **50**, (1983), pp. 346–349.

Gray, P. and Scott, S.K., Autocatalytic reactions in the isothermal continuous stirred tank reactor – Isolas and other forms of multistability, *Chem. Eng. Sci.*, **38**(1), (1983), pp. 29–43.

Gray, P. and Scott, S.K., Autocatalytic reactions in the isothermal, continuous stirred tank reactor – Oscillations and instabilities in the system $A + 2B \rightarrow 3B; B \rightarrow C$, *Chem. Eng. Sci.*, **39**(6), (1984), pp. 1087–1097.

Green, B.J., Wang, W., and Hudson J.L., Chaos and spatiotemporal pattern formation in electrochemical reactions, *Forma*, **15**, (2000), pp. 257 – 265.

Haykin, S, Neural Networks – A comprehensive foundation, *Macmillan*, (1994)

Hegger, R., Jaeger, L., and Kantz, H., Reconstruction of the dynamics of noisy multivariate time series, *preprint*, (1997). Available at: `http://www.mpiks-dresden.mpg.de/eprint/hegger/9702003/paper2.ps.gz`, 2002.

Hegger, R., Kantz, H., and Schreiber, T., Practical implementation of nonlinear time series methods: The TISEAN package., *CHAOS*, **9**, (1999), 413-435.

Hegger, H. and Schreiber, T., A noise reduction method for multivariate time series., *Physics Letters A*, **170**, (1992), pp. 305–310.

Hoffman, U and Schadlich, H.K., The influence of reaction orders and changes on the total number of moles on the conversion in a periodically operated CSTR, *Chem. Eng. Sci.*,(1986), **41**, (1986), pp. 2733–2738

Holcomb, G.R., Covino, B.S., Jr., and Eden, D., State-of-the-art review of electrochemical noise sensors, (2002). Available at: `http://www.netl.doe.gov/scng/publications/t&d/tsa/ENStateoftheArt.pdf`, 2002.

Hudson, J.L. and Tsotsis T.T., Electrochemical reaction dynamics: A Review, *Chem. Eng. Sci.*, **49**(10), (1994), pp. 1493 –1572.

Hunter, N.F., Application of nonlinear time time-series models to driven systems, *Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, Proc. Vol. XII*, eds. M. Casdagli and S. Eubank, (1992), pp. 467–491

Hyvärinen, A. and Oja, E., A fast fixed-point algorithm for independent component analysis, *Neural Computation*,**9**, (1997), pp. 1483–1492.

Hyvärinen, A. and Oja, E., Independent component analysis: Algorithms and Applications, *Neural Networks*,**13**(4–5), (2000), pp. 411–430.

Jollife, I.T., *Principal Component Analysis*, Springer-Verlag, New York, USA, (1986).

Jorgensen, D.V. and Aris, R., On the dynamics of a stirred tank with consecutive reactions, *Chem. Eng. Sci.*, **38**(1), (1983), pp. 45–53.

Judd, K., An improved estimator of dimension and some comments on providing confidence intervals, *Physica D*,**56**, (1992), pp. 216 – 228.

Judd, K. and Mees, A., On selecting models for nonlinear time series., *Physica D*, **82**, (1995), pp. 426–444.

Judd, K. and Mees A., Modeling chaotic motions of a string from experimental data., *Physica D*, **92**, (1996), pp. 221–236.

Judd, K. and Mees, K., Embedding as a modeling problem., *Physica D*, **120**, (1998), pp. 273–286.

Judd, K. and Small, M., Towards long-term prediction, *Physica D*, **136**, (2000), pp. 31–44.

Judd, K., Small, M., and Mees, A.I., Achieving good nonlinear models: Keep it simple, vary the embedding and get the dynamics right.,In: *Nonlinear Dynamics and Statistics*, ed. A.I. Mees, Birkhauser Boston, (2001), pp. 65 – 80.

Kaneko, K., Pattern dynamics in spatiotemporal chaos, *Physica D*, **34**, (1989), pp. 1–41.

Kaneko, K., Spatiotemporal chaos in one- and two-dimensional coupled map lattices, *Physica D*, **37**, (1989), pp. 60–82.

Kantz, H., and Schreiber, T., Nonlinear Time Series Analysis, *Cambridge University Press*, Cambridge, (1997).

Kennel, M.B., Brown, R., and Abarbanel, H.D.I., Determining embedding dimension for phase-space reconstruction using a geometrical construction, *Phys. Rev. A*, **45**, (1992), pp. 3403 – 3411.

Kennel, M.B. and Abarbanel, H.D.I., False neighbors and false strands: A reliable minimum embedding dimension algorithm, *Phys. Rev. E.*, **66**, (2002), pp. 1 – 18.

Kourti, T., Process Analysis and Abnormal Situation Detection: From Theory to Practice, *IEEE Control Sys. Mag.*, October 2002, pp. 10 – 25.

Kugiumtzis, D., Lillekjendlie, B., and Christophersen, N., Chaotic time series - Part I: Estimation of some invariant properties in state space, *Modeling, Identification and Control*, **15**(4), (1994), pp. 205-224.

Lazar, M.and Pastravanu, O., A neural predictive controller for non-linear systems, *Mathematics and Computers in Simulation*, **60**, (2002), pp. 315–324

Lee, J. and Chang, K., Applications of chaos and fractals in process systems engineering., *J. Proc. Cont.*, **6**, (1996), pp. 71–87.

Landa, P.S. and Rosenblum, M.G., Time series analysis for system identification and diagnostics, *Physica D*, **48**, (1991), 232–254.

Lillekjendlie, B., Kugiumtzis, N., and Christophersen, Chaotic time series Part II: System identification and prediction, *Modeling, Identification and Control*, **15**(4), (1994), pp. 225-243.

Lynch, D.T., Chaotic behavior of reaction systems: Parallel cubic autocatalators, *Chem. Eng. Sci.*, **47**(2), (1992), pp. 347 – 355

Lynch, D.T., Chaotic behavior of reaction systems: Mixed cubic and quadratic autocatalysis, *Chem. Eng. Sci.*, **47**(17/18), (1992), pp. 4435 – 4444

Mankin, J.C. and J.L. Hudson, Oscillatory and chaotic behaviour of a forced exothermic chemical reaction, *Chem. Eng. Sci.*, **39**(12), (1984), pp. 1807 –1814.

Mees, A.I., Rapp, P.E., and Jennings, L.S., Singular-value decomposition and embedding dimension, *Phys. Rev. A*, **36**(1), (1987), pp. 340 –346

Muldoon, M.R., Broomhead, D.S., Huke, J.P., and R. Hegger, Delay embedding in the presence of dynamical noise, *preprint*,(1998)

Müller, K.R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B., An introduction to kernel-based learning algorithms, *IEEE Trans. Neural Networks*, **12**(2), (2001), pp. 181 – 202

Ørstavik, S., Carretero-González, and Stark, J., Estimation of intensive quantities in spatio-temporal systems from time-series, *Physica D*, **147**, (2000), pp. 204–220.

Ørstavik, S and Stark, J., Reconstruction and cross-prediction in coupled map lattices using spatio-temporal embedding techniques, *Phys. Let. A*, **247**, (1998), pp. 145–160.

Packard, N.H., Crutchfield, J.P., Farmer, J.D., and Shaw, R.S., Geometry from a time series, *Phys. Rev. Lett.*, **45**, (1980), pp. 712–716

Paluš, M., Testing for nonlinearity using redundancies: quantitative and qualitative aspects, *Physica D*,**80**, (1995), pp. 186–205

Paluš, M., Detecting nonlinearity in multivariate time series, *Phys. Let. A*,**213**, (1996), pp. 138–147

Parekh, N., Kumar Ravi V., and Kulkarni, B.D., Analysis and characterization of complex spatio-temporal patterns in nonlinear reaction-diffusion systems, *Physica A*, **224**, pp. 369–381.

Porporato, A. and Ridolfi, L., Clues to the existence of deterministic chaos in river flow, *Int. J. Mod. Phys. B*, **10**(15), (1996), pp. 1821–1862.

Porporato, A. and Ridolfi, L, Multivariate nonlinear prediction of river flows, *J. Hydrology*, **248**, (2001), pp. 109–122.

Prichard, D. and Theiler, J., Generating surrogate data for time series with several simultaneously measured variables, *Phys. Rev. Let.*, **73**(7), (1994), pp. 951–954.

Prichard, D. and Theiler, J., Generalized redundancies for time series analysis, *Physica D*,**84**, (1995), pp. 476–493.

Rapp, P.E., Schmah, T.I., and Mees, A.I., Models of knowing and the investigation of dynamical systems, *Physica D*, **132**, (1999), pp. 133–149.

Rissanen, J., Hypothesis selection and testing by the MDL principle, *The Computer Journal*, **42**(4), (1999), pp. 260–269.

Rosenstein, M.T., Collins J.J. , and De Luca C.J., A practical method for calculating largest Lyapunov exponents from small data sets, *Physica D*, **65**, (1993), pp.117 – 134

Rosenstein, M.T., Collins J.J. , and De Luca C.J., Reconstruction expansion as a geometry-based framework for choosing proper delay times, *Physica D*, (1994), **73**, pp. 82 – 98.

Sauer, T., Yorke J.A., and Casdagli, M., Embeddelogy, *J. Stat. Phys.*, **65**, (1991), 579–616.

Schreiber, T., Constrained randomization of time series data, *Phys. Rev. Lett.*, **80**, (1998), pp. 2105–2108.

Schreiber, T., Interdisciplinary application of nonlinear time series methods, *Phys. Rep.*, **308**, (1999), pp. 1–64

Schreiber, T., Measuring information transfer, *Phys. Rev. Lett.*, **85**, (2000), pp. 461–464.

Schreiber, T. and Schmitz, A., Improved surrogate data for nonlinearity tests, *Phys. Rev. Lett*, **77**, (1996), pp. 635–639.

Schreiber, T., and Schimtz, A., Surrogate time series., *Physica D*, **142**, (2000), pp. 346–382.

Silverton, P.L., Hudgins, R.R., Adesina, A.A., Ross G.S., and Feimer, J.L., Activity and selectivity control through period composition forcing over Fischer-Tropsch catalysts, *Chem. Eng. Sci.*, **41**(4), (1986), pp. 923–928.

Small, M., Nonlinear dynamics in infant respiration., *PhD Thesis*, University of Western Australia, Dept. of Mathematics, (1998).

Small, M., Yu, D., and Harrison, R.G., Non-stationarity as an embedding problem, In: *Space Time Chaos: Characterization, Control, and Synchronization*, eds. S. Boccaletti *et al.*, World Scientific, (2001), pp. 3 – 18.

Small, M., and Judd, K., Comparisons of new nonlinear modeling techniques with application to infant respiration, *Physica D*, **117**, (1998), pp. 283–298.

Small, M. and Judd, K., Detecting nonlinearity in experimental data., *Int. J. Bifurcation and Chaos*, **8**, (1998), pp. 1231–1244.

Small, M., Judd, K., and Mees, A.I., Testing time series for nonlinearity., *Stat. and Comp.*, **11**, (2001), pp. 257–268.

Small, M., Yu, D., Simonotto, Harrison R.G., Grubb, N., and Fox, K.A.A., Uncovering non-linear structure in human ECG recordings., *Chaos, Sol., and Frac.*, **13**, (2002), pp. 1755–1762.

Smola A., Learning with kernels, *PhD Thesis*, Techische Universitäit Berlin, (1998).

Stark, J., Broomhead D.S., Davies, M.E., and Huke, J., Takens embedding theorems for forced and stochastic systems, *Nonlinear Analysis, Theory, Methods & Applications*, **30**(8),Proc. 2nd World Congress of Nonlinear Analysis, 1997, pp. 5303–5314.

Suykens, J.A.K, Nonlinear modelling and support vector machines, *IEEE Instrumentation and Measurement Technology Conference*, Budapest, Hungary, May 21 – 23, 2001.

Suykens, J.A.K., De Brabanter, J., Lucas, L., and Vandewalle, J., Weighted least squares support vector machines: robustness and sparse approximation, *Internal Report 00-37*, ESAT-SISTA, K.U.Leuven. Available at: `www.esat.kuleuven.ac.be\sista\lssvmlab`, 2002.

Takens, F, Detecting strange attractors in turbulence, in *Dynamical Systems and Turbulence*, eds. Rand, D.A. and Young, L.S., *Springer-Verlag*, Berlin, (1981), pp. 366–381.

Theiler, J., Galdrikian, B., Longtin, A, Eubank, S., and Farmer, J.D., Using surrogate data to detect nonlinearity in time series., *Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, Proc. Vol. XII*, eds. M. Casdagli and S. Eubank, Addison-Wesley, (1992), pp. 163–185.

Theiler, J., and Prichard, D., Constrained–realization Monte–Carlo method for hypothesis testing., *Physica D*, **94**, (1996), pp. 221–235.

Tong, H., Non-linear Time Series Analysis, Oxford University Press, Oxford, (1990).

van der Bleek, C.M., Schouten, J.C., and Coppens, MC., Chaos: Nourishment for new multiphase reactor development, paper presented at *The South African Institute of Chemical Engineers' SAIChE 2000 9th National Meeting*, 9 – 12 October 2000, Mpumalanga, South Africa.

Wagialla, K.M., Helal, A.M., and Elnashie, S.S.E.H., The use of the mathematical and computer models to explore the applicability of fluidized bed technology for highly exothermic reactions. 1. Oxidative dehydrogenation of butene, *Math. Comput. Modelling.*, **15**, pp. 27 – 31.

Wang, W., Kiss, I.Z., and Hudson, J.L., Clustering of Arrays of Chaotic Chemical Oscillators by Feedback and Forcing, *Phys. Rev. Lett.*, **86**(21), (2001), pp. 4954 - 4957.

Wiesenfeldt, M., Parlitz, U., and Lautorborn, W., Analysis of multi-channel data, *Proc. International workshop on advanced black-box techniques for nonlinear modeling*, Leuven, Belgium, (1998), pp. 138–146.

Wiesenfeldt, M., Parlitz, U., and Lautorborn, W., Mixed state analysis of multivariate time series, *Int. J. Bifurcation and Chaos*, **8**, (2001), 2217–2226.

Yu, D., Small, M., Harrison, R.G., and Diks, C., Efficient implementation of the Gaussian kernel algorithm in estimating invariants and noise level from noisy time series data, *Phys. Rev. E*, **61**, (2000), pp. 3750–3756.